

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7639258号

(P7639258)

(45)発行日 令和7年3月5日(2025.3.5)

(24)登録日 令和7年2月25日(2025.2.25)

(51)国際特許分類

F I

G 1 6 B 40/20 (2019.01)

G 1 6 B 40/20

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

1 3 0

G 1 6 B 15/30 (2019.01)

G 1 6 B 15/30

請求項の数 18 (全43頁)

(21)出願番号 特願2023-534153(P2023-534153)

(86)(22)出願日 令和4年3月9日(2022.3.9)

(65)公表番号 特表2023-552416(P2023-552416
A)

(43)公表日 令和5年12月15日(2023.12.15)

(86)国際出願番号 PCT/CN2022/079885

(87)国際公開番号 WO2022/206320

(87)国際公開日 令和4年10月6日(2022.10.6)

審査請求日 令和5年6月5日(2023.6.5)

(31)優先権主張番号 202110355929.6

(32)優先日 令和3年4月1日(2021.4.1)

(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者 514187420

テンセント・テクノロジー・(シェン
ジェン)・カンパニー・リミテッド
中華人民共和国 5 1 8 0 5 7 グアンド
ン, シェンジェン, ナンシャ・ディ
ストリクト, ミッドウエスト・ディスト
リクト・オブ・ハイテックパーク ケジ
ヨンギ・ロード テンセント・ビルディ
ング 3 5 エフ

(74)代理人 100107766

弁理士 伊東 忠重

(74)代理人 100070150

弁理士 伊東 忠彦

(74)代理人 100135079

弁理士 宮崎 修

最終頁に続く

(54)【発明の名称】 予測モデルの訓練方法、データ予測方法、装置及びコンピュータプログラム

(57)【特許請求の範囲】

【請求項1】

コンピュータ機器が実行する、予測モデルの訓練方法であって、

訓練サンプルセットを取得するステップであって、前記訓練サンプルセットは、各訓練サンプル、前記各訓練サンプルに対応する訓練サンプル重み、及び前記各訓練サンプルに対応する目標エネルギー特徴を含み、前記訓練サンプルは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、前記目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、前記野生型エネルギー特徴は、前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、前記変異型エネルギー特徴は、前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される、ステップと、

前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップと、

前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップと、

モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得するステ

ップであって、前記目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される、ステップと、を含む、

前記事前訓練予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とし、

前記基礎予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報及び前記訓練サンプル重みを出力とし、

前記目標予測モデルは、前記訓練サンプル及び前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とする、方法。

【請求項 2】

前記訓練サンプルセットを取得する前に、

前記各訓練サンプル及び前記各訓練サンプルに対応する相互作用状態ラベルを取得するステップであって、前記訓練サンプルは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含む、ステップと、

前記野生型タンパク質情報及び前記化合物情報に基づいて結合初期エネルギー特徴抽出を行い、野生型初期エネルギー特徴を取得するステップと、

前記変異型タンパク質情報及び前記化合物情報に基づいて結合初期エネルギー特徴抽出を行い、変異型初期エネルギー特徴を取得し、前記野生型初期エネルギー特徴及び前記変異型初期エネルギー特徴に基づいて前記各訓練サンプルに対応する目標初期エネルギー特徴を決定するステップと、

前記各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力して予測を行い、前記各訓練サンプルに対応する初期相互作用状態情報を取得するステップであって、前記初期予測モデルは、ランダムフォレストアルゴリズムを用いて構築される、ステップと、

前記各訓練サンプルに対応する初期相互作用状態情報及び前記各訓練サンプルに対応する相互作用状態ラベルに基づいて損失計算を行い、前記各訓練サンプルに対応する初期損失情報を取得するステップと、

事前訓練が完了するまで、前記初期損失情報に基づいて前記初期予測モデルを更新し、前記各訓練サンプルに対応する目標エネルギー特徴を初期予測モデルに入力して予測するステップを再度実行し、事前訓練予測モデル及び前記目標初期エネルギー特徴に対応する特徴重要度を取得するステップと、

事前訓練の完了時の前記各訓練サンプルに対応する損失情報に基づいて前記各訓練サンプルに対応する訓練サンプル重みを決定し、前記特徴重要度に基づいて前記目標初期エネルギー特徴から目標エネルギー特徴を選択するステップと、をさらに含む、請求項 1 に記載の方法。

【請求項 3】

前記各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力して予測を行い、前記各訓練サンプルに対応する初期相互作用状態情報を取得するステップであって、前記初期予測モデルは、ランダムフォレストアルゴリズムを用いて構築される、ステップは、

前記各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力するステップと、

前記初期予測モデルが前記各訓練サンプルに対応する目標初期エネルギー特徴を現在の分割すべきセットとし、前記目標初期エネルギー特徴に対応する初期特徴重要度を計算し、前記初期特徴重要度に基づいて前記目標初期エネルギー特徴から初期分割特徴を決定し、前記初期分割特徴に基づいて前記各訓練サンプルに対応する目標初期エネルギー特徴を分割し、各分割サンプルに対応する目標初期エネルギー特徴を含む各分割結果を取得し、分割が完了するまで、前記各分割結果を現在の分割すべきセットとし、前記目標初期エネルギー特徴に対応する初期特徴重要度を計算するステップを再度反復的に実行し、前記各訓練サンプルに対応する初期相互作用状態情報を取得するステップと、を含む、請求項 2 に記載の方法。

10

20

30

40

50

【請求項 4】

前記訓練サンプルセットを取得するステップであって、前記訓練サンプルセットは、前記各訓練サンプルに対応する訓練サンプル重みを含む、ステップは、

前記各訓練サンプルに対応する信頼度を取得し、前記信頼度に基づいて前記各訓練サンプルに対応する訓練サンプル重みを決定するステップ、を含む、請求項 1 に記載の方法。

【請求項 5】

前記訓練サンプルセットを取得するステップであって、前記訓練サンプルセットは、前記各訓練サンプルに対応する目標エネルギー特徴を含む、ステップは、

前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行い、前記野生型エネルギー特徴を取得するステップと、

前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行い、前記変異型エネルギー特徴を取得するステップと、

前記野生型エネルギー特徴と前記変異型エネルギー特徴との差を計算し、目標エネルギー特徴を取得するステップと、を含む、請求項 1 に記載の方法。

【請求項 6】

前記野生型エネルギー特徴は、第 1 の野生型エネルギー特徴及び第 2 の野生型エネルギー特徴を含み、

前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行い、前記野生型エネルギー特徴を取得するステップは、

前記野生型タンパク質情報及び前記化合物情報に基づいて非物理的なスコアリング関数を用いて結合エネルギー特徴抽出を行い、第 1 の野生型エネルギー特徴を取得するステップと、

前記野生型タンパク質情報及び前記化合物情報に基づいて物理的な関数を用いて結合エネルギー特徴抽出を行い、第 2 の野生型エネルギー特徴を取得するステップと、

前記第 1 の野生型エネルギー特徴及び前記第 2 の野生型エネルギー特徴に基づいて融合を行い、前記野生型エネルギー特徴を取得するステップと、を含む、請求項 5 に記載の方法。

【請求項 7】

前記変異型エネルギー特徴は、第 1 の変異型エネルギー特徴及び第 2 の変異型エネルギー特徴を含み、

前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行い、前記変異型エネルギー特徴を取得するステップは、

前記変異型タンパク質情報及び前記化合物情報に基づいて非物理的な関数を用いて結合エネルギー特徴抽出を行い、第 1 の変異型エネルギー特徴を取得するステップと、

前記変異型タンパク質情報及び前記化合物情報に基づいて物理的な関数を用いて結合エネルギー特徴抽出を行い、第 2 の変異型エネルギー特徴を取得するステップと、

前記第 1 の変異型エネルギー特徴及び前記第 2 の変異型エネルギー特徴に基づいて融合を行い、前記変異型エネルギー特徴を取得するステップと、を含む、請求項 5 に記載の方法。

【請求項 8】

前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップは、

タンパク質ファミリー情報を取得し、前記タンパク質ファミリー情報に基づいて前記訓練サンプルセットを分割し、各訓練サンプルグループを取得するステップと、

前記訓練サンプル重みに基づいて前記各訓練サンプルグループから現在の訓練サンプルを選択し、現在の訓練サンプルセットを取得するステップと、を含み、

前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップは、

前記現在の訓練サンプルセットにおける各現在の訓練サンプルに対応する現在の目標エ

10

20

30

40

50

エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、目標基礎予測モデルを取得するステップ、を含む、請求項 1 に記載の方法。

【請求項 9】

前記訓練サンプル重みに基づいて前記各訓練サンプルグループから現在の訓練サンプルを選択し、現在の訓練サンプルセットを取得するステップは、

現在の学習パラメータを取得し、前記現在の学習パラメータに基づいて選択サンプル数及びサンプル分布を決定するステップと、

前記選択サンプル数及び前記サンプル分布に基づいて前記訓練サンプル重みに従って前記各訓練サンプルグループから現在の訓練サンプルを選択し、目標の現在の訓練サンプルセットを取得するステップと、を含む、請求項 8 に記載の方法。

10

【請求項 10】

前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップは、

前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を前記事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得するステップと、

前記現在の相互作用状態情報と前記現在の訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、現在の損失情報を取得するステップと、

基礎訓練の完了条件に達するまで、前記現在の損失情報に基づいて前記事前訓練予測モデルを更新し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得するステップを再度実行し、前記基礎予測モデルを取得するステップと、を含む、請求項 1 に記載の方法。

20

【請求項 11】

前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新するステップは、

前記各訓練サンプルに対応する目標エネルギー特徴を前記基礎予測モデルに入力し、前記各訓練サンプルに対応する基礎相互作用状態情報を取得するステップと、

前記各訓練サンプルに対応する基礎相互作用状態情報と前記各訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、基礎損失情報を取得するステップと、

前記基礎損失情報に基づいて前記訓練サンプル重みを更新し、前記各訓練サンプルに対応する更新サンプル重みを取得するステップと、を含む、請求項 1 に記載の方法。

30

【請求項 12】

前記基礎損失情報に基づいて前記訓練サンプル重みを更新し、前記各訓練サンプルに対応する更新サンプル重みを取得するステップは、

現在の学習パラメータを取得し、前記現在の学習パラメータに基づいて更新閾値を計算するステップと、

前記更新閾値と前記各訓練サンプルに対応する基礎損失情報とを比較し、前記各訓練サンプルに対応する比較結果を取得するステップと、

前記各訓練サンプルに対応する比較結果に基づいて前記各訓練サンプルに対応する更新サンプル重みを決定するステップと、を含む、請求項 11 に記載の方法。

40

【請求項 13】

前記現在の学習パラメータは、多様性学習パラメータ及び難易度学習パラメータを含み、前記現在の学習パラメータに基づいて更新閾値を計算するステップは、

各訓練サンプルグループを取得し、前記各訓練サンプルグループから現在の訓練サンプルグループを決定し、前記現在の訓練サンプルグループに対応するサンプルランクを計算するステップと、

前記サンプルランクに基づいて加重値を計算し、前記加重値を用いて前記多様性学習パラメータに対して重み付けを行い、目標加重値を取得するステップと、

前記目標加重値と前記難易度学習パラメータとの和を計算し、前記更新閾値を取得するステップと、を含む、請求項 12 に記載の方法。

50

【請求項 14】

前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新した後に、

現在の学習パラメータを取得し、所定の増加量に従って前記現在の学習パラメータを更新し、更新学習パラメータを取得し、前記更新学習パラメータを現在の学習パラメータとするステップ、をさらに含む、請求項 1 に記載の方法。

【請求項 15】

データ予測方法であって、

予測すべきデータを取得するステップであって、前記予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む、ステップと、

前記予測すべき野生型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、前記予測すべき変異型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得するステップと、

前記予測すべき野生型エネルギー特徴及び前記予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定するステップと、

前記予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得するステップであって、前記目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定し、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される、ステップと、を含む、

前記事前訓練予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とし、

前記基礎予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報及び前記訓練サンプル重みを出力とし、

前記目標予測モデルは、前記訓練サンプル及び前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とする、方法。

【請求項 16】

予測モデルの訓練装置であって、

訓練サンプルセットを取得するサンプル取得モジュールであって、前記訓練サンプルセットは、各訓練サンプル、前記各訓練サンプルに対応する訓練サンプル重み、及び前記各訓練サンプルに対応する目標エネルギー特徴を含み、前記訓練サンプルは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、前記目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、前記野生型エネルギー特徴は、前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、前記変異型エネルギー特徴は、前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される、サンプル取得モジュールと、

前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するサンプル決定モジュールと、

前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得する訓練モジュールと、

モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得する反復

10

20

30

40

50

モジュールであって、前記目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される、反復モジュールと、を含み、

前記事前訓練予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とし、

前記基礎予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報及び前記訓練サンプル重みを出力とし、

前記目標予測モデルは、前記訓練サンプル及び前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とする、装置。

【請求項 17】

データ予測装置であって、

予測すべきデータを取得するデータ取得モジュールであって、前記予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む、データ取得モジュールと、

前記予測すべき野生型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、前記予測すべき変異型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得する特徴抽出モジュールと、

前記予測すべき野生型エネルギー特徴及び前記予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定する目標特徴決定モジュールと、

前記予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得する予測モジュールであって、前記目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定し、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される、予測モジュールと、を含み、

前記事前訓練予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とし、

前記基礎予測モデルは、前記目標エネルギー特徴を入力とし、前記相互作用状態情報及び前記訓練サンプル重みを出力とし、

前記目標予測モデルは、前記訓練サンプル及び前記目標エネルギー特徴を入力とし、前記相互作用状態情報を出力とする、装置。

【請求項 18】

コンピュータ読み取り可能な命令を含むコンピュータプログラムであって、前記コンピュータ読み取り可能な命令は、プロセッサにより実行される際に、請求項 1 乃至 15 の何れかに記載の方法のステップを実現する、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、2021年4月1日出願した出願番号が2021103559296であり、発明の名称が「予測モデルの訓練方法、データ予測方法、装置及び記憶媒体」である中国特許出願に基づく優先権を主張し、その全ての内容を参照により本発明に援用する。

【0002】

本発明の実施例は、コンピュータ技術の分野に関し、特に予測モデルの訓練方法、データ予測方法、装置、コンピュータ機器及び記憶媒体に関する。

【背景技術】

【0003】

人工知能技術の発展に伴い、化合物と標的タンパク質との親和性を予測するための機械

10

20

30

40

50

学習アルゴリズムの使用が出現した。現在、機械学習アルゴリズムを使用して構築されたモデルにより、標的タンパク質の変異後の化合物との親和性の変化を予測し、標的タンパク質が化合物に対して耐性を有するか否かを決定することによって、医師による薬の使用のために参照を提供する。しかし、現在の機械学習アルゴリズムにより構築された予測モデルは、正確率が低く、モデルの汎化能力が悪いという問題がある。

【発明の概要】

【0004】

このため、上記の技術的課題に対して、予測モデルの訓練正確性を向上させ、予測正確性を向上させることができる予測モデルの訓練方法、データ予測方法、装置、コンピュータ機器及び記憶媒体を提供する必要がある。

【0005】

1つの態様では、コンピュータ機器が実行する、予測モデルの訓練方法であって、訓練サンプルセットを取得するステップであって、前記訓練サンプルセットは、各訓練サンプル、前記各訓練サンプルに対応する訓練サンプル重み、及び前記各訓練サンプルに対応する目標エネルギー特徴を含み、前記訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、前記目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、前記野生型エネルギー特徴は、前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、前記変異型エネルギー特徴は、前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される、ステップと、前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップと、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップと、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得するステップであって、前記目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される、ステップと、を含む、方法を提供する。

【0006】

もう1つの態様では、予測モデルの訓練装置であって、訓練サンプルセットを取得するサンプル取得モジュールであって、前記訓練サンプルセットは、各訓練サンプル、前記各訓練サンプルに対応する訓練サンプル重み、及び前記各訓練サンプルに対応する目標エネルギー特徴を含み、前記訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、前記目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、前記野生型エネルギー特徴は、前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、前記変異型エネルギー特徴は、前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される、サンプル取得モジュールと、前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するサンプル決定モジュールと、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得する訓練モジュールと、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得する反復モジュールであって、前記目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される、反復モジュールと、を含む、装置を提供する。

【0007】

もう1つの態様では、コンピュータ読み取り可能な命令が記憶されたメモリと、プロセ

10

20

30

40

50

ッサと、を含むコンピュータ機器であって、前記プロセッサは、前記コンピュータ読み取り可能な命令を実行する際に、訓練サンプルセットを取得するステップであって、前記訓練サンプルセットは、各訓練サンプル、前記各訓練サンプルに対応する訓練サンプル重み、及び前記各訓練サンプルに対応する目標エネルギー特徴を含み、前記訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、前記目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、前記野生型エネルギー特徴は、前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、前記変異型エネルギー特徴は、前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される、ステップと、前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップと、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップと、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得するステップであって、前記目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される、ステップと、を実現する、コンピュータ機器を提供する。【0008】

もう1つの態様では、コンピュータ読み取り可能な命令が記憶されたコンピュータ読み取り可能な記憶媒体であって、前記コンピュータ読み取り可能な命令は、プロセッサにより実行される際に、訓練サンプルセットを取得するステップであって、前記訓練サンプルセットは、各訓練サンプル、前記各訓練サンプルに対応する訓練サンプル重み、及び前記各訓練サンプルに対応する目標エネルギー特徴を含み、前記訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、前記目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、前記野生型エネルギー特徴は、前記野生型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、前記変異型エネルギー特徴は、前記変異型タンパク質情報及び前記化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される、ステップと、前記訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップと、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップと、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得するステップであって、前記目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される、ステップと、を実現する、記憶媒体を提供する。【0009】

上記の予測モデルの訓練方法、装置、コンピュータ機器及び記憶媒体は、訓練サンプルセットを取得し、訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得し、目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される

。即ち、反復プロセスにおいて訓練サンプル重みを継続的に更新し、訓練サンプル重みを使用して訓練サンプルセットから現在の訓練サンプルを決定することによって、訓練サンプルの品質を確保し、その後、現在の訓練サンプルを使用して予測モデルを訓練することによって、訓練された目標予測モデルが予測の正確性及び汎化性を向上させることができる。

【0010】

もう1つの態様では、データ予測方法であって、予測すべきデータを取得するステップであって、前記予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む、ステップと、前記予測すべき野生型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、前記予測すべき変異型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得するステップと、前記予測すべき野生型エネルギー特徴及び前記予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定するステップと、前記予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得するステップであって、前記目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定し、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される、ステップと、を含む、方法を提供する。

【0011】

もう1つの態様では、データ予測装置であって、予測すべきデータを取得するデータ取得モジュールであって、前記予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む、データ取得モジュールと、前記予測すべき野生型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、前記予測すべき変異型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得する特徴抽出モジュールと、前記予測すべき野生型エネルギー特徴及び前記予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定する目標特徴決定モジュールと、前記予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得する予測モジュールであって、前記目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定し、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される、予測モジュールと、を含む、装置を提供する。

【0012】

もう1つの態様では、コンピュータ読み取り可能な命令が記憶されたメモリと、プロセッサと、を含むコンピュータ機器であって、前記プロセッサは、前記コンピュータ読み取り可能な命令を実行する際に、予測すべきデータを取得するステップであって、前記予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む、ステップと、前記予測すべき野生型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、前記予測すべき変異型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を

取得するステップと、前記予測すべき野生型エネルギー特徴及び前記予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定するステップと、前記予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得するステップであって、前記目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定し、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される、ステップと、を実現する、コンピュータ機器を提供する。

10

【0013】

もう1つの態様では、コンピュータ読み取り可能な命令が記憶されたコンピュータ読み取り可能な記憶媒体であって、前記コンピュータ読み取り可能な命令は、プロセッサにより実行される際に、予測すべきデータを取得するステップであって、前記予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む、ステップと、前記予測すべき野生型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、前記予測すべき変異型タンパク質情報及び前記予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得するステップと、前記予測すべき野生型エネルギー特徴及び前記予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定するステップと、前記予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得するステップであって、前記目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定し、前記現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、前記基礎予測モデルに基づいて前記訓練サンプル重みを更新し、訓練サンプル重みに基づいて前記訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される、ステップと、を実現する、記憶媒体を提供する。

20

30

【0014】

上記のデータ予測方法、装置、コンピュータ機器及び記憶媒体は、予測すべきデータを取得し、予測すべき目標エネルギー特徴を決定し、予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得する。目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される。即ち、目標予測モデルにより予測を行って相互作用状態情報を取得し、訓練により取得された目標予測モデルは、予測の正確性を向上させることができるため、取得された相互作用状態情報の正確性を向上させることができる。

40

【図面の簡単な説明】

【0015】

本発明の実施例に係る技術をより明確に説明するために、以下は、実施例の説明に必要な図面を簡単に紹介する。なお、以下の図面は、単に本発明の幾つかの実施形態であり、当業者は、創造的な作業を行うことなく、これらの図面に基づいて他の図面を得ることができる。

【図1】 1つの実施例に係る予測モデルの訓練方法の応用環境の図である。

50

【図 2】 1つの実施例に係る予測モデルの訓練方法の流れの概略図である。
【図 3】 1つの実施例に係る初期予測モデルの事前訓練の流れの概略図である。
【図 4】 1つの実施例に係る初期相互作用状態情報の取得の流れの概略図である。
【図 5】 1つの実施例に係る目標エネルギー特徴の取得の流れの概略図である。
【図 6】 1つの実施例に係る野生型エネルギー特徴の取得の流れの概略図である。
【図 7】 1つの実施例に係る変異エネルギー特徴の取得の流れの概略図である。
【図 8】 1つの実施例に係る目標基礎予測モデルの取得の流れの概略図である。
【図 9】 1つの実施例に係る基礎予測モデルの取得の流れの概略図である。
【図 10】 1つの実施例に係る更新サンプル重みの取得の流れの概略図である。
【図 11】 1つの実施例に係るデータ予測方法の流れの概略図である。
【図 12】 1つの具体的な実施例に係るデータ予測方法の応用シナリオの流れの概略図である。

10

【図 13】 1つの具体的な実施例に係る予測モデルの訓練方法の流れの概略図である。
【図 14】 1つの具体的な実施例に係る予測モデルの訓練方法の流れの概略図である。
【図 15】 1つの具体的な実施例に係る対比試験結果の概略図である。
【図 16】 図 15 の具体的な実施例に係る適合率と再現率の曲線の指標の概略図である。
【図 17】 1つの実施例に係る予測モデルの訓練装置の構成のブロック図である。
【図 18】 1つの実施例に係るデータ予測装置の構成のブロック図である。
【図 19】 1つの実施例に係るコンピュータ機器の内部構成図である。
【図 20】 もう 1つの実施例に係るコンピュータ機器の内部構成図である。
【発明を実施するための形態】

20

【0016】

本発明の目的、技術的手段、及び利点をより明確にするために、以下は、図面を参照しながら本発明の実施形態をさらに詳細に説明する。なお、本明細書に記載される具体的な実施例は、単に本発明を説明するためのものであり、本発明を限定するものではない。

【0017】

本発明に係る予測モデルの訓練方法は、図 1 に示す応用環境に適用されてもよい。ここで、端末 102 はネットワークを介してサーバ 104 と通信する。サーバ 104 は端末 102 により送信されたモデル訓練命令を受信する。サーバ 104 は、モデル訓練命令に基づいてデータベース 106 から訓練サンプルセットを取得する。訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、野生型エネルギー特徴は、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、変異型エネルギー特徴は、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される。サーバ 104 は、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定する。サーバ 104 は、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得する。サーバ 104 は、モデル訓練が完了するまで、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得する。目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される。ここで、端末 102 は、各種のパーソナルコンピュータ、ラップトップコンピュータ、スマートフォン、タブレットコンピュータ、及び携帯型ウェアラブルデバイスであってもよいが、これらに限定されない。サーバ 104 は、独立したサーバ又は複数のサーバからなるサーバクラスターを用いて実現されてもよい。

30

40

【0018】

1つの実施例では、図 2 に示すように、予測モデルの訓練方法を提供する。該方法は、

50

図1のサーバに適用されることを一例にして説明するが、該方法は、端末に適用されてもよいし、端末とサーバを含むシステムに適用されてもよく、端末とサーバとのインタラクションにより実現される。本実施例では、該方法は、以下のステップを含む。

【0019】

ステップ202において、訓練サンプルセットを取得する。訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、野生型エネルギー特徴は、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、変異型エネルギー特徴は、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される。

10

【0020】

ここで、タンパク質とは、タンパク質キナーゼなどの標的タンパク質を意味する。化合物とは、標的タンパク質と相互作用可能な薬物である。例えば、チロシンキナーゼ阻害薬である。タンパク質情報は、標的タンパク質の具体的な情報を表すために使用され、タンパク質構造、タンパク質の物理化学的性質などを含んでもよい。野生型タンパク質情報とは、自然から得られた個体、即ち非人工的に変異誘発されたタンパク質に関する情報を意味する。変異型タンパク質情報とは、例えば、薬物構造の変異などの変異が生じたタンパク質に関する情報を意味する。化合物情報とは、タンパク質と相互作用可能な化合物の情報を意味し、化合物の構造、化合物の物理化学的性質などを含んでもよい。訓練サンプル重みは、対応する訓練サンプルの品質を表すために使用される、訓練サンプルに対応する重みを意味する。高品質の訓練サンプルは、機械学習モデルを訓練する際に訓練の品質を向上させることができる。結合エネルギー特徴とは、標的タンパク質と化合物分子との相互作用エネルギー情報を表すために使用される、タンパク質と化合物との相互作用時の特徴を意味し、構造的特徴、物理化学的特性、及び特性選択により得られた特徴であるエネルギー特徴などを含んでもよい。野生型エネルギー特徴とは、野生型タンパク質と化合物とが相互作用する際に抽出された結合エネルギー特徴を意味する。変異型エネルギー特徴とは、変異型タンパク質と化合物が相互作用する際に抽出された結合エネルギー特徴を意味する。目標エネルギー特徴は、変異型エネルギー特徴と野生型エネルギー特徴との差異を表すために使用される。

20

30

【0021】

具体的には、サーバは、データベースから訓練サンプルセットを直接取得してもよい。訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、野生型エネルギー特徴は、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、変異型エネルギー特徴は、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される。サーバは、インターネットから各訓練サンプルを収集し、各訓練サンプルに対応する目標エネルギー特徴を抽出し、各訓練サンプルに対応する訓練サンプル重みを初期化してもよい。サーバは、データサービスを提供するサードパーティのサーバから訓練サンプルセットを取得してもよく、例えば、サードパーティのクラウドサーバから訓練サンプルセットを取得してもよい。

40

【0022】

1つの実施例では、サーバは、タンパク質情報、変異型タンパク質情報、及び化合物情報を取得し、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行って野生型エネルギー特徴を取得し、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行って変異型エネルギー特徴を取得し、野生型エネルギー特徴

50

と変異型エネルギー特徴との差を計算して目標エネルギー特徴を取得してもよい。また、初期化するに対応する訓練サンプル重みは、例えば、ランダム初期化、ゼロ初期化、ガウス分布初期化などであってもよい。

【0023】

ステップ204において、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定する。

【0024】

ここで、現在の訓練サンプルとは、現在の訓練で使用される訓練サンプルである。

【0025】

具体的には、サーバは、各訓練サンプルに対応する訓練サンプル重みに基づいて、訓練サンプルセットから訓練サンプルの選択を行い、現在の訓練サンプルを取得する。例えば、訓練サンプル重みが所定の重み閾値よりも大きい訓練サンプルを、現在の訓練サンプルとしてもよい。1つの具体的な実施例では、訓練サンプル重みは、0及び1に設定されてもよく、即ち、各訓練サンプルに対応する訓練サンプル重みは、0又は1に初期化されてもよい。訓練サンプル重みが1である場合、対応する訓練サンプルは、現在の訓練サンプルである。1つの実施例では、サーバは、複数の訓練サンプルを含む現在の訓練サンプルセットを取得するために、訓練サンプル重みに基づいて、訓練サンプルセットから複数の訓練サンプルを選択してもよい。現在の訓練サンプルセットを用いて基礎予測モデルの訓練を行う。

【0026】

ステップ206において、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得する。

【0027】

ここで、現在の目標エネルギー特徴とは、現在の訓練サンプルに対応する目標エネルギー特徴を意味する。事前訓練予測モデルとは、事前訓練された予測モデルを意味し、該予測モデルは、ランダムフォレストアルゴリズムを使用して構築され、変異前後のタンパク質と化合物との親和性の変化を予測するために使用されてもよい。基礎予測モデルは、訓練サンプル重みを保持したまま、対応する現在の訓練サンプルを用いて訓練することで得られる。

【0028】

具体的には、サーバは、事前訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測を行い、予測結果を取得し、該予測結果に基づいて損失を計算し、損失に基づいて事前訓練予測モデルを逆に更新し、基礎訓練完了条件に達するまで、事前訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測を行うステップを繰り返し実行し、基礎訓練完了条件に達した予測モデルを基礎予測モデルとしてもよい。ここで、基礎訓練完了条件とは、基礎予測モデルを取得するための条件を意味し、訓練が所定の反復回数の上限に達すること、又は損失が所定の閾値に達すること、又はモデルのパラメータが変化しなくなることなどを含む。

【0029】

ステップ208において、モデルの訓練が完了したか否かを判断し、モデルの訓練が完了した場合、ステップ208aを実行し、モデルの訓練が完了していない場合、ステップ208bを実行し、ステップ204に戻って実行する。

【0030】

ステップ208aにおいて、目標予測モデルを取得する。目標予測モデルは、入力されたタンパク質情報と入力された化合物情報とに対応する相互作用状態情報を予測するために使用される。

【0031】

ステップ208bは、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプ

10

20

30

40

50

ルを決定するステップの実行に戻る。

【0032】

ここで、モデル訓練完了とは、目標予測モデルが得られる条件であり、目標予測モデルとは、入力されたタンパク質情報と入力された化合物情報とに対応する相互作用状態情報を予測するために最終的に訓練されたモデルである。相互作用状態情報を用いて、タンパク質変異の前後と化合物間の結合自由エネルギーの変化を表す。結合自由エネルギーとは、リガンドと受容体の間に存在する相互作用を意味する。

【0033】

具体的には、サーバは、基礎予測モデルを取得した際に、モデル訓練完了に達したか否かをさらに判断し、該モデル訓練完了条件は、反復回数が予め設定されたモデル訓練反復回数10の上限に達することを含んでもよい。モデル訓練完了条件に達していない場合、基礎予測モデルのパラメータを変更せずに、基礎予測モデルを使用して各訓練サンプルに対応する訓練サンプル重みを更新し、各訓練サンプルに対応する目標エネルギー特徴を基礎予測モデルに入力し、各訓練サンプルに対応する損失を取得し、各訓練サンプルに対応する損失に基づいて各訓練サンプルに対応する訓練サンプル重みを更新する。訓練サンプル重みが更新されると、モデル訓練完了条件が達成されるまで、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを反復的に実行し続け、モデル訓練完了条件が達成された時点の基礎予測モデルを目標予測モデルとし、該目標予測モデルは、入力された化合物情報に対応する入力されたタンパク質情報の相互作用状態情報を予測するために使用される。20

【0034】

上記の予測モデルの訓練方法は、訓練サンプルセットを取得し、訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得し、目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される。即ち、反復プロセスにおいて訓練サンプル重みを継続的に更新し、訓練サンプル重みを使用して訓練サンプルセットから現在の訓練サンプルを決定することによって、訓練サンプルの品質を確保し、その後、現在の訓練サンプルを使用して予測モデルを訓練することによって、訓練された目標予測モデルが予測の正確性及び汎化性を向上させることができる。30

【0035】

1つの実施例では、図3に示すように、ステップ202の前に、即ち、訓練サンプルセットを取得する前に、以下のステップをさらに含む。

【0036】

ステップ302において、各訓練サンプル及び各訓練サンプルに対応する相互作用状態ラベルを取得する。訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含む。40

【0037】

ステップ304において、野生型タンパク質情報及び前記化合物情報に基づいて結合初期エネルギー特徴抽出を行い、野生型初期エネルギー特徴を取得する。

【0038】

ここで、結合初期エネルギー特徴とは、抽出されたスクリーニングされていない特徴を意味し、非物理的モデル特徴、物理経験的ポテンシャルに基づく特徴などを含んでもよい。ここで、非物理的モデル特徴は、結晶性タンパク質化合物構造特徴、リガンド及び残基50

の物理化学的特性、経験的又は記述子スコアリング関数に基づいて計算されたエネルギー特徴などを含む。物理と経験ポテンシャルエネルギーに基づく特徴とは、混合の物理と経験ポテンシャルエネルギーに基づくモデリングプログラムにより計算されたエネルギー特徴を意味する。野生型初期エネルギー特徴とは、野生型タンパク質情報と化合物情報を相互作用する時に抽出された結合初期エネルギー特徴を意味する。

【0039】

具体的には、サーバは、データベースから各訓練サンプルを取得してもよく、各訓練サンプルは事前訓練に使用されるサンプルであってもよい。各訓練サンプルは、訓練サンプルセット内の訓練サンプルと同一であってもよいし、異なってもよい。サーバは、インターネットから各訓練サンプルを収集してもよく、サーバは、データサービスを提供するサーバから各訓練サンプルを取得してもよい。各訓練サンプルは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含む。この場合、サーバは、各訓練サンプルについて特徴抽出を行ってもよく、即ち、野生型タンパク質情報及び化合物情報を用いて結合初期エネルギー特徴抽出を行い、各訓練サンプルに対応する野生型初期エネルギー特徴を取得する。

10

【0040】

ステップ306において、変異型タンパク質情報及び化合物情報に基づいて結合初期エネルギー特徴抽出を行い、変異型初期エネルギー特徴を取得し、野生型初期エネルギー特徴及び変異型初期エネルギー特徴に基づいて各訓練サンプルに対応する目標初期エネルギー特徴を決定する。

20

【0041】

ここで、変異型初期エネルギー特徴とは、変異型タンパク質情報と化合物情報との相互作用の際に抽出された結合初期エネルギー特徴を意味し、目標初期エネルギー特徴は、野生型初期エネルギー特徴と変異型初期エネルギー特徴との差異を表すために使用される。

【0042】

具体的には、サーバは、変異型タンパク質情報及び化合物情報に対して結合初期エネルギー特徴抽出を行い、変異型初期エネルギー特徴を取得し、野生型初期エネルギー特徴と変異型初期エネルギー特徴との差異を計算し、該差異を目標初期エネルギー特徴とする。例えば、構造特徴間の差異を計算し、該差異を目標構造特徴としてもよい。或いは、物理化学的性質間の差異を計算し、物理化学的性質間の差異を目標構造特徴としてもよい。

30

【0043】

ステップ308において、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力して予測を行い、各訓練サンプルに対応する初期相互作用状態情報を取得する。初期予測モデルは、ランダムフォレストアルゴリズムを用いて構築される。

【0044】

ここで、初期予測モデルとは、モデルパラメータ初期化の予測モデルを意味し、このモデルパラメータ初期化は、ランダム初期化であってもよいし、ゼロ初期化であってもよい。初期予測モデルは、ランダムフォレストアルゴリズムを使用して構築され、ランダムフォレストとは、複数の木を使用してサンプルを訓練して予測する分類器を意味する。例えば、ExtraTree（極端ランダム木）アルゴリズムを使用して初期予測モデルを構築してもよい。初期相互作用状態情報とは、初期予測モデルを用いて予測された相互作用状態情報を意味する。

40

【0045】

具体的には、サーバは、予めランダムフォレストアルゴリズムを用いてモデルパラメータ初期化のための初期予測モデルを作成し、そして、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力して予測し、出力された各訓練サンプルに対応する初期相互作用状態情報を取得する。

【0046】

ステップ310において、各訓練サンプルに対応する初期相互作用状態情報及び各訓練サンプルに対応する相互作用状態ラベルに基づいて損失計算を行い、各訓練サンプルに対

50

応する初期損失情報を取得する。

【0047】

ここで、相互作用状態ラベルとは、真の相互作用状態情報を意味し、訓練サンプルごとに対応する相互作用状態ラベルがある。初期損失情報は、初期相互作用状態情報と相互作用状態ラベルとの誤差を表すために使用される。

【0048】

具体的には、サーバは、予め設定された損失関数を用いて、各訓練サンプルに対応する初期相互作用状態情報と相互作用状態ラベルとの損失を計算し、各訓練サンプルに対応する初期損失情報を取得する。ここで、損失関数は、平均二乗誤差損失関数、平均絶対値誤差損失関数などであってもよい。

10

【0049】

ステップ312において、事前訓練が完了するまで、初期損失情報に基づいて初期予測モデルを更新し、各訓練サンプルに対応する目標エネルギー特徴を初期予測モデルに入力して予測するステップを再度実行し、事前訓練予測モデル及び目標初期エネルギー特徴に対応する特徴重要度を取得する。

【0050】

ここで、事前訓練完了とは、事前訓練予測モデルを取得する条件であり、事前訓練回数が予め設定された反復回数に達したこと、又は事前訓練の損失が予め設定された閾値に達したこと、又は事前訓練の予測モデルパラメータが変化しなくなったことを意味する。特徴重要度は、目標初期エネルギー特徴の重要レベルを表すために用いられ、特徴重要度が高いほど、対応する特徴が重要となり、モデル訓練時の寄与が大きくなる。

20

【0051】

具体的には、サーバは、初期損失情報を使用して勾配を計算し、その後、勾配を使用して初期予測モデルを逆更新し、更新された予測モデルを取得し、事前訓練が完了したか否かを判断し、事前訓練が完了していない場合、事前訓練が完了するまで、更新された予測モデルを初期予測モデルとし、各訓練サンプルに対応する目標エネルギー特徴を初期予測モデルに入力して予測するステップを再度反復的に実行し、最後の反復で得られた更新された予測モデルを事前訓練予測モデルとする。事前訓練予測モデルは、ランダムフォレストアルゴリズムを使用して構築されるため、事前訓練予測モデルを訓練完了する場合、目標初期エネルギー特徴に対応する特徴重要度を直接得ることができる。目標初期エネルギー特徴の各特徴は、対応する特徴重要度を有する。

30

【0052】

ステップ314において、事前訓練の完了時の各訓練サンプルに対応する損失情報に基づいて各訓練サンプルに対応する訓練サンプル重みを決定し、特徴重要度に基づいて目標初期エネルギー特徴から目標エネルギー特徴を選択する。

【0053】

具体的には、サーバは、事前訓練が完了したときの各訓練サンプルに対応する損失情報を用いて、各訓練サンプルに対応する訓練サンプル重みを決定してもよい。例えば、各訓練サンプルに対応する損失情報と重み損失閾値とを比較し、損失情報が重み損失閾値よりも大きい場合、対応する訓練サンプルは品質の良いサンプルであり、対応する訓練サンプル重みを1に設定してもよい。損失情報が重み損失閾値以下である場合、対応する訓練サンプルは品質の悪いサンプルとなり、対応する訓練サンプル重みは0に設定してもよい。特徴重要度により目標初期エネルギー特徴から特徴選択を行い、目標エネルギー特徴を取得し、目標エネルギー特徴は、事前訓練予測モデルが更に訓練される時に取得する特徴である。

40

【0054】

上記の実施例では、各訓練サンプルを用いて事前訓練して事前訓練モデルを取得した後、事前訓練完了時の各訓練サンプルに対応する損失情報に基づいて各訓練サンプルに対応する訓練サンプル重みを決定し、特徴重要度に基づいて目標初期エネルギー特徴から特徴選択を行って目標エネルギー特徴を取得することによって、更なる訓練時に訓練効率を向

50

上させ、訓練の正確性を保証することができる。

【0055】

1つの実施例では、図4に示すように、ステップ308、即ち、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力して予測を行い、各訓練サンプルに対応する初期相互作用状態情報を取得するステップであって、初期予測モデルは、ランダムフォレストアルゴリズムを用いて構築される、ステップは、以下のステップを含む。

【0056】

ステップ402において、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力する。

【0057】

ステップ404において、初期予測モデルが各訓練サンプルに対応する目標初期エネルギー特徴を現在の分割すべきセットとし、目標初期エネルギー特徴に対応する初期特徴重要度を計算し、初期特徴重要度に基づいて目標初期エネルギー特徴から初期分割特徴を決定し、初期分割特徴に基づいて各訓練サンプルに対応する目標初期エネルギー特徴を分割し、各分割サンプルに対応する目標初期エネルギー特徴を含む各分割結果を取得し、分割が完了するまで、各分割結果を現在の分割すべきセットとし、目標初期エネルギー特徴に対応する初期特徴重要度を計算するステップを再度反復的に実行し、各訓練サンプルに対応する初期相互作用状態情報を取得する。

【0058】

ここで、初期特徴重要度とは、目標初期エネルギー特徴に対応する特徴重要度であり、初期分割特徴とは、決定木分割を行う特徴である。分割結果とは、目標初期エネルギー特徴を分割したものであり、分割サンプルとは、分割結果における目標初期エネルギー特徴に対応する訓練サンプルである。

【0059】

具体的には、サーバは、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力し、初期予測モデルは入力特徴を採点し、目標初期エネルギー特徴に対応する初期特徴重要度を取得する。ここで、情報利得、情報利得率、ジニ係数、平均二乗誤差などを用いて初期特徴重要度を計算してもよい。初期特徴重要度に基づいて目標初期エネルギー特徴から初期分割特徴を決定し、初期分割特徴に基づいて各訓練サンプルに対応する目標初期エネルギー特徴を分割し、即ち、初期分割特徴を超える目標初期エネルギー特徴を一部とし、初期分割特徴を超えない目標初期エネルギー特徴を他の部分とし、分割結果を取得する。分割結果は、各分割サンプルが対応する目標初期エネルギー特徴を含む。分割が完了するまで、各分割結果を現在の分割すべきセットとし、目標初期エネルギー特徴に対応する初期特徴重要度を計算するステップを反復的に実行し、各訓練サンプルに対応する初期相互作用状態情報を取得する。ここで、分割が完了するとは、各樹ノードを分割できないことを意味し、即ち、葉ノードが唯一の目標初期エネルギー特徴にしか対応していない。初期相互作用状態情報とは、初期予測モデルにより予測された相互作用状態情報である。

【0060】

上記の実施例では、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力し、初期予測モデルが目標初期エネルギー特徴に対応する初期特徴重要度を計算し、初期特徴重要度に基づいて目標初期エネルギー特徴から初期分割特徴を決定し、初期分割特徴に基づいて各訓練サンプルに対応する目標初期エネルギー特徴を分割し、各分割サンプルに対応する目標初期エネルギー特徴を含む各分割結果を取得し、分割が完了するまで、各分割結果を現在の分割すべきセットとし、目標初期エネルギー特徴に対応する初期特徴重要度を計算するステップを再度反復的に実行し、各訓練サンプルに対応する初期相互作用状態情報を取得することによって、得られた初期相互作用状態情報の正確性を高めることができる。

【0061】

1つの実施例では、ステップ202、即ち、訓練サンプルセットを取得するステップで

10

20

30

40

50

あって、訓練サンプルセットは、前記各訓練サンプルに対応する訓練サンプル重みを含む、ステップは、以下のステップを含む。

【0062】

各訓練サンプルに対応する信頼度を取得し、信頼度に基づいて各訓練サンプルに対応する訓練サンプル重みを決定する。

【0063】

ここで、信頼度は、訓練サンプルの品質の良さのレベルを表すために用いられる。信頼度が高いほど、訓練サンプルに対応する品質が高くなり、信頼度の高い訓練サンプルを用いて訓練されたモデルの性能が良くなる。

【0064】

具体的には、サーバは、各訓練サンプルを取得する際に、各訓練サンプルに対応する信頼度を同時に取得してもよい。そして、信頼度の間の差をそのまま各訓練サンプルに対する訓練サンプル重みとしてもよい。ここで、この信頼度は、人為的に設定されたものであってもよいし、予め各訓練サンプルについて信頼度評価を行って得られたものであってもよい。1つの実施例では、各訓練サンプルに対応する信頼度と予め設定された信頼度閾値とを比較し、信頼度閾値を超える場合、対応する訓練サンプルの重みを1に設定し、この訓練サンプルを現在の訓練サンプルとしてもよい。信頼閾値を超えていない場合は、対応する訓練サンプル重みを0に設定する。

【0065】

上記の実施例では、信頼度を取得し、信頼度に基づいて各訓練サンプルに対応する訓練サンプル重みを決定することによって、訓練サンプル重みの取得効率を向上させることができる。

【0066】

1つの実施例では、図5に示すように、ステップ202、即ち、訓練サンプルセットを取得するステップであって、訓練サンプルセットは、各訓練サンプルに対応する目標エネルギー特徴を含む、ステップは、以下のステップを含む。

【0067】

ステップ502において、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行い、野生型エネルギー特徴を取得する。

【0068】

ステップ504において、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行い、変異型エネルギー特徴を取得する。

【0069】

ここで、野生型エネルギー特徴は、野生型タンパク質特徴、化合物特徴、及び野生型タンパク質情報と化合物情報とが相互作用するときのエネルギー特徴を含むが、これらに限定されない。野生型タンパク質特徴は、野生型タンパク質情報に対応する特徴を表すために使用され、野生型タンパク質構造特徴、野生型タンパク質物理化学性質特徴を含むが、これらに限定されない。化合物の特徴は、化合物の構造特徴、化合物の物理化学的性質特徴を含むが、これらに限定されない。変異型エネルギー特徴は、変異型タンパク質特徴、化合物特徴、及び変異型タンパク質情報と化合物情報とが相互作用する場合のエネルギー特徴を含むが、これらに限定されない。変異タンパク質特徴は、変異タンパク質情報に対応する特徴を表すために使用され、変異タンパク質構造特徴、変異タンパク質の物理化学的性質特徴を含むが、これらに限定されない。

【0070】

具体的には、サーバは、野生型タンパク質情報及び化合物情報を用いて特徴抽出を行い、野生型タンパク質特徴及び化合物特徴に抽出するとともに、野生型タンパク質と化合物とが相互作用する際のエネルギー特徴を抽出し、野生型タンパク質特徴、化合物特徴及びエネルギー特徴を野生型エネルギー特徴とする。サーバは、変異型タンパク質情報を用いて特徴抽出を行い、変異型タンパク質特徴をした後、変異型タンパク質と化合物とが相互作用する際のエネルギー特徴を抽出し、抽出された変異型タンパク質特徴化合物特徴及び

10

20

30

40

50

エネルギー特徴を変異型エネルギー特徴とする。

【0071】

ステップ506において、野生型エネルギー特徴と変異型エネルギー特徴との差を計算し、目標エネルギー特徴を取得する。

【0072】

具体的には、サーバは、野生型のエネルギー特徴と変異型のエネルギー特徴との間の差を計算し、例えば、野生型のタンパク質特徴と変異型のタンパク質特徴との差を計算し、野生型のタンパク質と化合物とが相互作用する場合のエネルギー特徴と変異型のタンパク質と化合物とが相互作用する場合のエネルギー特徴との差を計算し、目標エネルギー特徴を取得する。1つの具体的な態様では、野生型エネルギー特徴と変異型エネルギー特徴との特徴差を計算し、目標エネルギー特徴を取得してもよい。

10

【0073】

上記の態様では、野生型エネルギー特徴と変異型エネルギー特徴とを抽出し、野生型エネルギー特徴と変異型エネルギー特徴との差を算出し、目標エネルギー特徴を取得することによって、得られた目標エネルギー特徴の正確性を向上させることができる。

【0074】

1つの実施例では、野生型エネルギー特徴は、第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴を含む。

【0075】

図6に示すように、ステップ502、即ち、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行い、野生型エネルギー特徴を取得するステップは、以下のステップを含む。

20

【0076】

ステップ602において、野生型タンパク質情報及び化合物情報に基づいて非物理的なスコアリング関数を用いて結合エネルギー特徴抽出を行い、第1の野生型エネルギー特徴を取得する。

【0077】

ここで、非物理的なスコアリング関数とは、経験又は記述子に基づくスコアリング関数を意味し、該スコアリング関数は、幾つかの事前の仮定に基づいて、或いは実験データを適合させて、明らかに説明可能な物理的意味を持たないエネルギー特徴を取得する。第1の野生型エネルギー特徴は、抽出された第1の部分エネルギー特徴を意味する。

30

【0078】

具体的には、サーバは、予め設定された非物理的なスコアリング関数を用いて結合エネルギー特徴抽出を行い、野生型タンパク質情報及び化合物情報を非物理型スコアリング関数により計算し、計算結果を取得し、計算結果を第1の野生型エネルギー特徴としてもよい。ここで、スコアリング関数（理論的に取得された受容体ーリガンド結合モードの合理性を評価する関数）を用いてエネルギー特徴を抽出してもよい。

【0079】

ステップ604において、野生型タンパク質情報及び化合物情報に基づいて物理的な関数を用いて結合エネルギー特徴抽出を行い、第2の野生型エネルギー特徴を取得する。

40

【0080】

ここで、物理的な関数は、混合の物理と経験ポテンシャルエネルギーに基づくエネルギー関数を意味し、明確な物理意義を持ち、エネルギー関数ファミリーは実験データのフィッティングに基づく力場関数、第一性原理に基づく量化計算関数、連続媒質に基づく溶剤モデルなどから構成される。

【0081】

具体的には、サーバは、予め設定された物理的な関数を用いて野生型タンパク質情報と化合物信に対して結合エネルギー特徴抽出を行い、第2の野生型エネルギー特徴を取得する。例えば、エネルギー特徴は、混合された物理的及び経験的ポテンシャルエネルギーに基づくモデリングプログラムRosetta（モンテカルロシミュレーションアニーリン

50

グをアルゴリズムの核心とする高分子モデリングのソフトウェアライブラリ)におけるエネルギー関数を用いて計算されてもよい。

【0082】

ステップ606において、第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴に基づいて融合を行い、野生型エネルギー特徴を取得する。

【0083】

具体的には、サーバは、第1の野生型エネルギー特徴と第2の野生型エネルギー特徴との特徴差を計算し、野生型エネルギー特徴を取得する。

【0084】

上記の実施例では、第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴を抽出し、第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴に基づいて融合を行い、野生型エネルギー特徴を取得する。第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴は、野生型標的タンパク質と化合物分子との相互作用エネルギー情報をよりよく表すことができるため、得られる野生型エネルギー特徴をより正確にすることができる。

10

【0085】

1つの実施例では、変異型エネルギー特徴は、第1の変異型エネルギー特徴及び第2の変異型エネルギー特徴を含む。

【0086】

図7に示すように、ステップ504、即ち、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行い、変異型エネルギー特徴を取得するステップは、以下のステップを含む。

20

【0087】

ステップ702において、変異型タンパク質情報及び化合物情報に基づいて非物理的な関数を用いて結合エネルギー特徴抽出を行い、第1の変異型エネルギー特徴を取得する。

【0088】

ステップ704において、変異型タンパク質情報及び化合物情報に基づいて物理的な関数を用いて結合エネルギー特徴抽出を行い、第2の変異型エネルギー特徴を取得する。

【0089】

具体的には、サーバは、予め設定された非物理的な関数を用いて変異型タンパク質情報と化合物情報との結合エネルギー特徴抽出を行い、第1の変異型エネルギー特徴を取得した後、予め設定された物理的な関数を用いて変異型タンパク質情報と化合物情報との結合エネルギー特徴抽出を行い、第2の変異型エネルギー特徴を取得する。

30

【0090】

ステップ706において、第1の変異型エネルギー特徴及び第2の変異型エネルギー特徴に基づいて融合を行い、変異型エネルギー特徴を取得する。

【0091】

具体的には、サーバは、第1の変異型エネルギー特徴と第2の変異型エネルギー特徴との間の特徴差を計算し、変異型エネルギー特徴を取得する。

【0092】

上記の実施例では、第1の変異型エネルギー特徴及び第2の変異型エネルギー特徴を抽出し、第1の変異型エネルギー特徴及び第2の変異型エネルギー特徴に基づいて融合を行い、変異型エネルギー特徴を取得する。第1の変異型エネルギー特徴と第2の変異型エネルギー特徴とが変異型標的タンパク質と化合物分子との間の相互作用エネルギー情報をよりよく表すことができるため、得られた変異型エネルギー特徴がより正確になる。

40

【0093】

1つの実施例では、図8に示すように、ステップ204、即ち、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップは、以下のステップを含む。

【0094】

50

ステップ 802 において、タンパク質ファミリー情報を取得し、タンパク質ファミリー情報に基づいて訓練サンプルセットを分割し、各訓練サンプルグループを取得する。

【0095】

ここで、生体内でアミノ酸配列が類似し、且つ構造や機能が十分に類似したタンパク質が「タンパク質ファミリー」(protein family)を構成しており、同一のタンパク質ファミリーのメンバーを「相同タンパク質」と称される。タンパク質ファミリー情報はタンパク質ファミリーの情報であり、訓練サンプルセットは同一のタンパク質ファミリーに対応する各訓練サンプルを分割して得られる。

【0096】

具体的には、サーバは、データベースからタンパク質ファミリー情報を直接取得してもよく、このタンパク質ファミリー情報は、インターネットから取得されてもよいし、データサービスを提供する第三者サーバから取得されてもよい。1つの実施例では、サーバは、訓練サンプルにおけるタンパク質情報の構造又は配列が類似しているタンパク質ファミリーを同一の訓練サンプルグループに分割して、各訓練サンプルグループを取得してもよい。

10

【0097】

ステップ 804 において、訓練サンプル重みに基づいて各訓練サンプルグループから現在の訓練サンプルを選択し、現在の訓練サンプルセットを取得する。

【0098】

具体的には、サーバは、訓練サンプル重みを使用して、各訓練サンプルグループから現在の訓練サンプルを選択し、即ち、訓練サンプルグループ内の訓練サンプル重みに従って現在の訓練サンプルを順次選択するとともに、各訓練サンプルグループから選択して、現在の訓練サンプルセットを取得する。

20

【0099】

ステップ 206、即ち、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップは、以下のステップを含む。

【0100】

ステップ 806 において、現在の訓練サンプルセットにおける各現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、目標基礎予測モデルを取得する。

30

【0101】

具体的には、サーバは、基礎訓練のために、現在の訓練サンプルセットにおける各現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力し、基礎訓練が完了すると、目標基礎予測モデルを取得する。

【0102】

上記の実施例では、タンパク質ファミリー情報に基づいて前記訓練サンプルセットを分割し、各訓練サンプルグループを取得し、そして、訓練サンプル重みに基づいて各訓練サンプルグループから現在の訓練サンプルを選択し、現在の訓練サンプルセットを取得し、現在の訓練サンプルセットにおける各現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎予測モデルを取得する。即ち、各訓練サンプルグループから現在の訓練サンプルを選択することによって、選択された訓練サンプルを1つの局部領域に集中することではなく、空間全体に分布させるため、モデルを訓練する際に、訓練サンプルに含まれる大域情報を学習できることを保証し、モデルの訓練過程における学習知識の全面性を保証し、さらにモデル訓練過程における収束速度を向上させると共に、訓練によって得られたモデルの汎化性を向上させることができる。

40

【0103】

1つの具体的な実施例では、事前訓練予測モデルの基本的な形式は、式(1)で示される。

50

【 0 1 0 4】

【数 1】

$$\min_{w \in \Gamma, v \in [0,1]^n} \sum_{i=1}^n v_i L(g_w(x_i), y_i) - \lambda \|v\|_1 - \gamma \|v\|_{2,1} \quad \text{式 (1)}$$

ここで、 n は訓練サンプルの総数を表し、 X は訓練サンプル集合を表し、 $X = (x_1, \dots, x_n) \in \mathbb{R}^n \times m$ 、 $\mathbb{R}$ は実数集合を表し、 m はエネルギー特徴の数を表す。 x_i は i 番目の訓練サンプルを表し、 y_i は i 番目の訓練サンプルに対応する相互作用状態ラベルを表す。 g は事前訓練予測モデルを表し、 w はモデルパラメータを表し、 L は損失関数を表し、 v は訓練サンプル重みを表す。 $v = (v^{(1)}, \dots, v^{(b)})$ 、 b は訓練サンプルグループのグループ数を表し、即ち訓練サンプルセットを b 個のグループ、即ち $x^{(1)}, \dots, x^{(b)}$ に分割する。ここで、

(外 1)

$$X^{(j)} \in \mathbb{R}^{n_j \times m}$$

は j 番目の訓練サンプルグループの訓練サンプルを表し、

(外 2)

$$v^{(j)} = (v_1^j, \dots, v_{n_j}^j) \in [0, 1]^{n_j}$$

、 n_j は j 番目の訓練サンプルグループにおける訓練サンプル数を表し、

(外 3)

$$\sum_{j=1}^b n_j = n, \quad v_1^j$$

、

(外 4)

$$v_1^j$$

は j 番目の訓練サンプルグループにおける 1 番目の訓練サンプルに対する訓練サンプル重みを表す。 v_i は i 番目の訓練サンプル重みを表す。 λ は訓練サンプルの難易度のパラメータを表し、即ち、訓練サンプルを選択する際に、選択しやすい（信頼性が高い）サンプルから選択しにくい（信頼性が低い）サンプルへと順次選択されることを表す。 γ はサンプル多様性のパラメータを表す。即ち、複数の訓練サンプルグループからサンプルを選択することを意味する。

(外 5)

$$\| \cdot \|_1$$

は L_1 ノルムを表し、
(外 6)

$$\| \cdot \|_{2,1}$$

10

は $L_{2,1}$ ノルムを表す。ここで、
(外 7)

$$\|V\|_{2,1} = \sum_{j=1}^b \|V^{(j)}\|_2$$

20

、 b は訓練サンプルグループのグループ数を表し、 j は j 番目の訓練サンプルグループの訓練サンプル重みを表す。即ち、負の L_1 ノルムは、信頼性の高いサンプル、即ち訓練の結果の誤差が小さいサンプルを選択する傾向がある。負の $L_{2,1}$ ノルムは、複数の訓練サンプルグループから訓練サンプルを選択し、多様性情報を予測モデルに埋め込むのに有利である。

【0105】

1つの実施例では、訓練サンプル重みに基づいて各訓練サンプルグループから現在の訓練サンプルを選択し、現在の訓練サンプルセットを取得するステップは、以下のステップを含む。

30

【0106】

現在の学習パラメータを取得し、現在の学習パラメータに基づいて選択サンプル数及びサンプル分布を決定し、選択サンプル数及びサンプル分布に基づいて訓練サンプル重みに従って各訓練サンプルグループから現在の訓練サンプルを選択し、目標の現在の訓練サンプルセットを取得する。

【0107】

ここで、現在の学習パラメータとは、現在の訓練で使用される学習パラメータであり、該現在の学習パラメータは、現在の訓練サンプルの選択を制御するために使用される。選択サンプル数とは、現在選択されている訓練サンプルの数である。サンプル分布とは、選択された現在の訓練サンプルの各訓練サンプルグループ内の分布である。目標の現在の訓練サンプルセットとは、現在の学習パラメータを使用して選択された現在の訓練サンプルのセットである。

40

【0108】

具体的には、サーバは、現在の訓練サンプルパラメータを取得し、該現在の訓練サンプルパラメータの初期値は予め設定されていてもよい。サーバは、現在の学習パラメータを使用して、現在訓練時に選択されるサンプル数とサンプル分布を計算する。そして、選択サンプル数及びサンプル分布に基づいて訓練サンプル重みに従って各訓練サンプルグループから現在の訓練サンプルを選択し、目標の現在の訓練サンプルセットを取得する。

【0109】

上記の実施例では、現在の学習パラメータを使用して訓練サンプルの選択をさらに制御

50

し、目標の現在の訓練サンプルセットを取得することによって、選択された訓練サンプルをより正確にすることができるため、訓練によって得られた予測モデルをより正確にし、予測モデルの汎化能力を向上させることができる。

【0110】

1つの実施例では、図9に示すように、ステップ206、即ち、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得するステップは、以下のステップを含む。

【0111】

ステップ902において、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得する。

10

【0112】

ここで、現在の相互作用状態情報を用いて、予測により得られた現在の訓練サンプル中の変異前後のタンパク質と化合物の相互作用の変化を表す。

【0113】

具体的には、サーバは、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルの入力として直接使用し、事前訓練予測モデルは、入力された現在の目標エネルギー特徴に基づいて予測し、予測結果である現在の相互作用状態情報を出力する。

【0114】

ステップ904において、現在の相互作用状態情報と現在の訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、現在の損失情報を取得する。

20

【0115】

ここで、現在の損失情報は、現在の訓練サンプルに対応する予測結果と真の結果との誤差である。

【0116】

具体的には、サーバは、現在の訓練サンプルに対応する相互作用状態ラベルを取得し、この相互作用状態ラベルは、予め設定されていてもよい。相互作用状態ラベルは、実験によって測定された変異前後のタンパク質と化合物の相互作用の変化であってもよい。そして、サーバは、予め設定された損失関数を用いて、現在の相互作用状態情報と現在の訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、現在の損失情報を取得する。

【0117】

ステップ906において、基礎訓練の完了条件に達するまで、現在の損失情報に基づいて事前訓練予測モデルを更新し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得するステップを再度実行し、基礎予測モデルを取得する。

30

【0118】

具体的には、サーバは、予め設定された基礎訓練反復回数に達するまで、又はモデルパラメータが変化しなくなるまで、現在の損失情報を使用して勾配降下アルゴリズムにより事前訓練予測モデルのパラメータを逆更新し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得するステップを再度反復的に実行し、最後の反復の事前訓練予測モデルを基礎予測モデルとする。

40

【0119】

1つの具体的な実施例では、事前訓練予測モデルに対応する最適化関数は、式(2)に示すように、回帰最適化関数である。

【0120】

【数2】

$$\min_{\mathcal{W} \in \Gamma} \sum_{i=1}^n v_i^* L(\mathcal{G}_{\mathcal{W}}(X_i), Y_i) \quad \text{式 (2)}$$

ここで、 v_i は訓練サンプル重みが重み閾値を超える訓練サンプルを選択して訓練することを表す。例えば、訓練サンプル重みが 0 と 1 のみを含む場合、訓練サンプル重みが 1 の訓練サンプルを選択して訓練してもよい。

【0121】

上記の実施例では、訓練サンプル重みを一定に維持した後、現在の訓練サンプルを選択して事前訓練予測モデルを訓練して、基礎予測モデルを取得することによって、訓練された基礎予測モデルをより正確にすることができる。

【0122】

1つの実施例では、図10に示すように、ステップ208b、即ち、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新するステップは、以下のステップを含む。

【0123】

ステップ1002において、各訓練サンプルに対応する目標エネルギー特徴を基礎予測モデルに入力し、各訓練サンプルに対応する基礎相互作用状態情報を取得する。

【0124】

ここで、各訓練サンプルとは、訓練サンプルセット内の各訓練サンプルである。基礎相互作用状態情報とは、基礎予測モデルにより予測された訓練サンプルごとに対応する相互作用状態情報である。この相互作用状態情報は、野生型タンパク質及び化合物の結合自由エネルギーと、変異型タンパク質及び化合物の結合自由エネルギーとの相対差であってもよい。

【0125】

具体的には、サーバは、訓練により基礎予測モデルを取得する際に、基礎予測モデル中のパラメータを不変に保持し、訓練サンプルセット中の各訓練サンプルに対応する訓練サンプル重みを更新する。即ち、サーバは、各訓練サンプルに対応する目標エネルギー特徴を基礎予測モデルに入力し、出力された各訓練サンプルに対応する基礎相互作用状態情報を取得する。

【0126】

ステップ1004において、各訓練サンプルに対応する基礎相互作用状態情報と各訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、基礎損失情報を取得する。

【0127】

ここで、基礎損失情報とは、基礎予測モデルの予測結果と実際の結果との誤差を意味する。

【0128】

具体的には、サーバは、予め設定された損失関数を使用して、各訓練サンプルの誤差、即ち、基礎相互作用状態情報と相互作用状態ラベルとの誤差を計算し、各訓練サンプルに対応する基礎損失情報を取得する。

【0129】

ステップ1006において、基礎損失情報に基づいて訓練サンプル重みを更新し、各訓練サンプルに対応する更新サンプル重みを取得する。

【0130】

具体的には、サーバは、各訓練サンプルに対応する基礎損失情報を使用して各訓練サンプル重みを更新し、サーバは、各訓練サンプルに対応する基礎損失情報を、各訓練サンプルに対応する更新サンプル重みとして直接使用してもよい。

【0131】

10

20

30

40

50

1つの実施例では、ステップ1006、即ち、基礎損失情報に基づいて訓練サンプル重みを更新し、各訓練サンプルに対応する更新サンプル重みを取得するステップは、以下のステップを含む。

【0132】

現在の学習パラメータを取得し、現在の学習パラメータに基づいて更新閾値を計算し、更新閾値と各訓練サンプルに対応する基礎損失情報とを比較し、各訓練サンプルに対応する比較結果を取得し、各訓練サンプルに対応する比較結果に基づいて各訓練サンプルに対応する更新サンプル重みを決定する。

【0133】

ここで、更新閾値とは、訓練サンプルの重みを更新するための閾値である。

10

【0134】

具体的には、サーバは、現在の学習パラメータを取得し、現在の学習パラメータを用いて更新閾値を決定する。更新閾値と各訓練サンプルに対応する基礎損失情報とを比較する。基礎損失情報が更新閾値を超えた場合、該訓練サンプルに対応する予測誤差が大きいことを意味し、この場合、対応する訓練サンプル重みを第1の訓練サンプル重みに更新する。基礎損失情報が更新閾値を超えていない場合、誤差が小さいことを意味し、この場合、対応する訓練サンプル重みを第2の訓練サンプル重みに更新する。そして、現在の訓練サンプルを選択する際に、第2の訓練サンプルの重みに対応する訓練サンプルを現在の訓練サンプルとして選択する。

【0135】

20

1つの実施例では、現在の学習パラメータは、多様性学習パラメータ及び難易度学習パラメータを含む。現在の学習パラメータに基づいて更新閾値を計算するステップは、以下のステップを含む。

【0136】

各訓練サンプルグループを取得し、各訓練サンプルグループから現在の訓練サンプルグループを決定し、現在の訓練サンプルグループに対応するサンプルランクを計算する。サンプルランクに基づいて加重値を計算し、加重値を用いて多様性学習パラメータに対して重み付けを行い、目標加重値を取得する。目標加重値と難易度学習パラメータとの和を計算し、更新閾値を取得する。

【0137】

30

ここで、難易度学習パラメータとは、難易度を評価する学習パラメータであり、難易度学習パラメータは、訓練時に選択された訓練サンプルの信頼度を決定するために使用される。多様性学習パラメータは、多様性を評価する学習パラメータである。多様性学習パラメータは、訓練時に選択された訓練サンプルの訓練サンプルセットにおける分布を決定するために使用される。サンプルランクは、事前訓練サンプルグループ中の訓練サンプルのランクであり、一つのベクトルグループのランクは、その最大の無関係のグループに含まれるベクトルの数である。現在の訓練サンプルグループとは、現在訓練サンプルの重みを更新する必要がある訓練サンプルグループを意味する。

【0138】

40

具体的には、サーバは、各訓練サンプルセットを取得し、各訓練サンプルセットから現在の訓練サンプルセットを決定し、現在の訓練サンプルセットに対応するサンプルランクを計算する。サンプルランクに基づいて重みを計算し、重みを用いて多様性学習パラメータを重み付けし、目標重みを取得する。目標重み値と難易度学習パラメータとの和を計算し、現在の訓練サンプルグループに対する更新閾値を取得する。1つの具体的な実施例では、各訓練サンプルグループ内の訓練サンプルは、基礎損失情報に従って昇順に並べ替えられる。それぞれの順序付けられた訓練サンプルグループを取得し、順序付けられた訓練サンプルグループ中で現在の訓練サンプルグループを決定し、現在の訓練サンプル組に対応する更新閾値を計算する。

【0139】

1つの具体的な実施例では、以下に示す式(3)を使用して、訓練サンプルに対応する

50

訓練サンプル重みを更新してもよい。

【 0 1 4 0 】

【 数 3 】

$$V_i^{(i)} = \begin{cases} 1, L\left(g_{w^*}^{(j)}\left(X_i^{(j)}\right), y_i^j\right) < \lambda + \gamma \frac{1}{\sqrt{a} + \sqrt{a-1}} \\ 0, L\left(g_{w^*}^{(j)}\left(X_i^{(j)}\right), y_i^j\right) \geq \lambda + \gamma \frac{1}{\sqrt{a} + \sqrt{a-1}} \end{cases} \quad \text{式 (3)} \quad 10$$

ここで、 a は j 番目の訓練サンプルセットにおけるランクを表す。
(外 8)

$$g_{w^*}^{(j)}\left(X_i^{(j)}\right)$$

20

は j 番目の訓練サンプルグループの i 番目の訓練サンプルに対応する予測された相互作用状態情報を表し、
(外 9)

$$y_i^j$$

は j 番目の訓練サンプルグループの i 番目の訓練サンプルに対応する真の相互作用状態ラベルを表す。
(外 10)

30

$$\lambda + \gamma \frac{1}{\sqrt{a} + \sqrt{a-1}}$$

は計算された更新閾値を表す。 j 番目の訓練サンプルグループの i 番目の訓練サンプルに対応する誤差が更新閾値よりも小さい場合、対応する訓練サンプル重みを 1 に更新し、 j 番目の訓練サンプルグループの i 番目の訓練サンプルに対応する誤差が更新閾値以上である場合、対応する訓練サンプル重みを 0 に更新する。

40

【 0 1 4 1 】

上記の実施例では、サンプル重みを常に更新し、訓練のために現在の訓練サンプルを再選択することによって、訓練プロセスにおいて誤差の大きい訓練サンプルを使用して訓練することができるため、誤差の大きい訓練サンプルが訓練プロセスに悪影響を及ぼすことを回避し、訓練によって得られた目標予測モデルの正確性を向上させることができる。

【 0 1 4 2 】

1つの実施例では、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新した後に、以下のステップを含む。

【 0 1 4 3 】

50

現在の学習パラメータを取得し、所定の増加量に従って現在の学習パラメータを更新し、更新学習パラメータを取得し、更新学習パラメータを現在の学習パラメータとする。

【0144】

具体的には、サーバは、現在の学習パラメータの更新条件を予め設定してもよく、例えば、重み更新毎の現在の学習パラメータの増加量を予め設定してもよい。そして、所定の増加量により現在の学習パラメータを更新し、更新学習パラメータを取得し、更新学習パラメータを現在の学習パラメータとする。1つの実施例では、サーバは、予め設定された増加するサンプル数を取得し、予め設定された増加するサンプル数によって現在の学習パラメータを更新し、更新された学習パラメータを取得し、更新された学習パラメータを現在の学習パラメータとしてもよい。そして、サンプル数が増加した後、訓練によって得られた損失情報が小さいものから大きくなった場合、訓練が完了し、サンプル数が増加しなかった場合に訓練によって得られた予測モデルを最終的に得られた目標予測モデルとする。

10

【0145】

1つの実施例では、図11に示すように、データ予測方法を提供する。該方法は、図1のサーバに適用されることを一例に説明するが、該方法は、端末に適用されてもよいし、端末とサーバを含むシステムに適用されてもよく、端末とサーバとのインタラクションによって実現される。本実施例では、該方法は、以下のステップを含む。

【0146】

ステップ1102において、予測すべきデータを取得する。予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む。

20

【0147】

ここで、予測すべき野生型タンパク質情報とは、相互作用状態情報を予測する必要がある野生型タンパク質情報である。予測すべき変異型タンパク質情報とは、相互作用状態の情報を予測する必要がある変異型タンパク質情報である。予測化合物情報とは、相互作用状態情報を予測する必要がある化合物情報である。

【0148】

具体的には、サーバは、インターネットから予測すべきデータを取得してもよいし、端末から予測すべきデータを取得してもよい。サーバは、データベースから予測データを直接取得してもよい。1つの実施例では、サーバは、サードパーティのサーバによって送信された予測データを取得してもよい。サードパーティのサーバは、ビジネスサービスを提供するサーバであってもよい。予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む。1つの実施例では、サーバは、予測すべき変異型タンパク質情報及び予測すべき化合物情報を端末から取得し、次いで、予測すべき変異型タンパク質情報に対応する予測すべき野生型タンパク質情報をデータベースから取得して、予測すべきデータを取得してもよい。

30

【0149】

ステップ1104において、予測すべき野生型タンパク質情報及び予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、予測すべき変異型タンパク質情報及び予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得する。

40

【0150】

ここで、予測すべき野生型エネルギー特徴とは、抽出された予測すべき野生型タンパク質情報と予測すべき化合物情報とが相互作用するときのエネルギー特徴を意味する。予測すべき変異型エネルギー特徴とは、予測すべき変異型タンパク質情報と予測すべき化合物情報とが相互作用するときに抽出されたエネルギー特徴である。

【0151】

具体的には、サーバは、予測すべき野生型タンパク質情報及び予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得する。例えば、予測すべき野生型タンパク質情報におけるタンパク質構造及び予測すべき化合物

50

情報における化合物構造に基づいて構造特徴を抽出し、次いで、予測すべき野生型タンパク質情報における物理化学的特性及び予測すべき化合物情報における物理化学的特性に基づいて物理化学的特性を抽出してもよい。物理化学的性質は、溶融沸点、常温での状態、色、酸性度を含む化学的性質などを含む化学物質の特性を測る指標である。予測すべき野生型タンパク質情報と予測すべき化合物情報との相互作用のエネルギー特徴は、スコアリング関数を用いて計算され、エネルギー特徴は、混合された物理的及び経験的ポテンシャルエネルギーに基づくエネルギー関数を用いて計算され、予測すべき野生型エネルギー特徴が得られる。次いで、予測すべき変異型タンパク質情報及び予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得し、例えば、予測すべき変異型タンパク質情報のタンパク質構造及び予測すべき化合物情報の化合物構造に基づいて構造特徴を抽出し、次いで、予測すべき変異型タンパク質情報の物理化学的性質及び予測すべき化合物情報の物理化学的性質に基づいて物理化学的性質特徴を抽出し、スコアリング関数を使用してエネルギー特徴を抽出し、物理的及び経験的ポテンシャルエネルギーに基づくエネルギー関数を使用してエネルギー特徴を抽出することによって、予測すべき変異型エネルギー特徴を取得してもよい。

10

【0152】

ステップ1106において、予測すべき野生型エネルギー特徴及び予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定する。

【0153】

具体的には、サーバは、予測すべき野生型エネルギー特徴の各特徴値と、予測すべき変異型エネルギー特徴に対応する特徴値との差を計算して、予測すべき目標エネルギー特徴を取得する。

20

【0154】

ステップ1108において、予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得する。目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される。

30

【0155】

ここで、目標予測モデルは、上記の予測モデルの訓練方法の何れかの実施例で訓練されたモデルであってもよい。即ち、目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される。

40

【0156】

具体的には、サーバは、予測すべき目標エネルギー特徴を目標予測モデルに入力して予測し、出力された相互作用状態情報を取得する。1つの具体的な実施例では、相互作用状態情報は、予測すべき変異型タンパク質及び予測すべき野生型タンパク質のそれぞれの、予測すべき化合物に対する結合自由エネルギーの相対的な差を意味する。その後、結合自由エネルギーの相対的な差と耐性閾値とを比較し、結合自由エネルギーの相対的な差が耐性閾値を超える場合、予測すべき変異タンパク質が耐性を持ち、使用を継続できなくなることを意味する。結合自由エネルギーの相対差が耐性閾値を超えない場合、予測すべき変異タンパク質は耐性を持たず、依然として正常に使用できることを意味する。

【0157】

50

上記のデータ予測方法、装置、コンピュータ機器及び記憶媒体は、予測すべきデータを取得し、予測すべき目標エネルギー特徴を決定し、予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得する。目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される。即ち、目標予測モデルにより予測を行って相互作用状態情報を取得し、訓練により取得された目標予測モデルは、予測の正確性を向上させることができるため、取得された相互作用状態情報の正確性を向上させることができる。

【0158】

本発明は、上記のデータ予測方法を適用する応用シナリオをさらに提供する。図12は、1つの具体的な実施例に係るデータ予測方法の応用シナリオの流れの概略図である。具体的には、標的タンパク質の変異が薬剤耐性を引き起こすことを予測する適用シナリオにおいて、サーバは、野生型タンパク質情報及び変異型タンパク質情報、並びに化合物情報を含む2つの異なるタイプの標的タンパク質情報を含む、端末に送信された予測データを取得する。次いで、野生型タンパク質情報及び変異型タンパク質情報、並びに化合物情報を使用して、非物理的モデルの特徴及び物理的及び経験的ポテンシャルに基づく特徴を含む、タンパク質変異後の親和性を予測するための参照価値を有する特徴を抽出する。結晶性タンパク質ーリガンド構造、リガンド及び残基の物理化学的特性、並びに経験的又は記述子のスコアリング関数に基づいて計算された幾つかのエネルギー特徴などの非物理的モデルの特性は、物理的及び経験的ポテンシャルに基づく特性に基づいて、混合物理的及び経験的ポテンシャルに基づくモデリングプログラムRosettaを使用して計算されたエネルギー特徴である。そして、特徴選択を行い、即ち、訓練時の特徴選択を経て得られた目標エネルギー特徴によって抽出された特徴から対応する特徴を選択し、予測すべき目標エネルギー特徴を選択し、予測すべき目標エネルギー特徴を目標予測モデルに入力して予測し、予測された結合自由エネルギーの差を取得する。この結合自由エネルギーの差と薬品の耐性閾値とを比較し、結合自由エネルギーの差が耐性閾値を超える場合、このタンパク質変異は耐薬品性を引き起こすタンパク質変異であることを意味する。結合自由エネルギーの差が耐薬品性閾値を超えない場合、このタンパク質変異は耐薬品性を引き起こさないタンパク質変異であることを意味する。この場合、予測結果を端末に送信して表示する。

【0159】

1つの具体的な実施例では、予測モデルの訓練方法を提供する。図13に示すように、該方法は、具体的に以下のステップを含む。

【0160】

ステップ1302において、訓練サンプルセットを取得する。訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、野生型エネルギー特徴は、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、変異型エネルギー特徴は、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される。

【0161】

ステップ1304において、タンパク質ファミリー情報を取得し、タンパク質ファミリー情報に基づいて訓練サンプルセットを分割し、各訓練サンプルグループを取得し、現在の学習パラメータを取得し、現在の学習パラメータに基づいて選択サンプル数及びサン

ル分布を決定する。サンプル数及びサンプル分布に基づいて訓練サンプル重みに従って各訓練サンプルグループから現在の訓練サンプルを選択し、目標の現在の訓練サンプルセットを取得する。

【0162】

ステップ1306において、目標の現在の訓練サンプルセットにおける各訓練サンプルに対応する目標エネルギー特徴を基礎予測モデルに入力し、各訓練サンプルに対応する基礎相互作用状態情報を取得し、各訓練サンプルに対応する基礎相互作用状態情報と各訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、基礎損失情報を取得する。

【0163】

ステップ1308において、現在の訓練サンプルグループに対応するサンプルランクを計算する。サンプルランクに基づいて加重値を計算し、加重値を用いて多様性学習パラメータに対して重み付けを行い、目標加重値を取得し、目標加重値と難易度学習パラメータとの和を計算し、各訓練サンプルグループの更新閾値を取得する。

10

【0164】

ステップ1310において、更新閾値と各訓練サンプルグループにおける訓練サンプルに対応する基礎損失情報とを比較し、各訓練サンプルに対応する比較結果を取得し、各訓練サンプルに対応する比較結果に基づいて各訓練サンプルグループにおける各訓練サンプルに対応する更新サンプル重みを決定する。

【0165】

ステップ1312において、所定の増加量に従って現在の学習パラメータを更新し、更新学習パラメータを取得し、更新学習パラメータを現在の学習パラメータとし、モデル訓練が完了するまで、現在の学習パラメータに基づいて選択サンプル数及びサンプル分布を決定するステップを再度実行し、目標予測モデルを取得する。

20

【0166】

本発明は、上記の予測モデル訓練方法を適用する応用シナリオをさらに提供する。具体的には、以下の通りである。

【0167】

図14は、1つの具体的な実施例に係る予測モデルの訓練方法の流れの概略図である。具体的には、該方法は、以下のステップを含む。

【0168】

入力データ及び訓練サンプルグループ情報を取得する。該入力データは、各訓練サンプルを含み、対応する訓練サンプル重みは0又は1である。該訓練サンプルグループ情報は、入力データ内の訓練サンプルが属する訓練サンプルグループを示す。この際に、予測モデルのモデルパラメータと学習パラメータを初期化する。

30

【0169】

その後、訓練サンプルに対応する訓練サンプル重みを固定したまま、モデルのパラメータを訓練する。即ち、初期化された学習パラメータに基づいて訓練サンプルの重みが1である訓練サンプルを選択し、現在の訓練サンプルを取得し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を抽出し、現在の目標エネルギー特徴を初期化された予測モデルに入力して基礎訓練を行い、基礎訓練が完了すると、基礎予測モデルを取得する。

40

【0170】

次に、基礎予測モデルのパラメータを固定したまま、サンプル重みを更新する、即ち、式(3)を用いて訓練サンプルごとに訓練サンプル重みを更新し、更新サンプル重みを取得する。

【0171】

この際に、初期化された学習パラメータをさらに更新し、モデル訓練が完了するまで、訓練サンプルに対応する訓練サンプル重みを固定したまま、モデルのパラメータを訓練するステップを再度反復的に実行し、訓練完了時の予測モデルのモデルパラメータ及び訓練サンプル重みを出力し、即ち、目標予測モデルを取得する。

【0172】

50

該実施例では、訓練された目標予測モデルに対して対比試験を行った。具体的には、訓練及びテストは、耐性基準データセットPlatinum（Platinumは、リガンドとプロテオームの相互作用に対するミスセンス変異の影響を研究し、理解するために開発された、薬剤耐性に関する情報を広く集めたデータベースである）及びTKIを使用して実施される。ここで、目標予測モデルは、データセットPlatinumを使用して訓練され、次に、データセットTKIを使用してテストされる。予測すべきタンパク質変異後の親和性変化に対応する特徴は、RDKit（RDKitは、ケモインフォマティクスのためのオープンソースのツールキットであり、化合物2Dと3Dの分子操作に基づいて、機械学習法を利用して化合物ディスクリプタ生成、fingerprint生成、化合物構造類似性計算、2Dと3D分子表示などを行う）、Biopython（Biopythonは、バイオインフォマティクスを使用及び研究する開発者のためのオンラインのリポジトリを提供する）、FoldX（タンパク質結合自由エネルギーを計算する）、PLIP（タンパク質リガンドの非共有結合相互作用の分析ツールである）、AutoDock（オープンソースの分子シミュレーションソフトウェアであり、最も主にリガンドー蛋白分子ドッキングの実行に応用されている）などの非物理的モデルツールを採用することで生成された。そして、混合された物理的及び経験的ポテンシャルエネルギーに基づくモデリングプログラムRosettaを用いてエネルギー特徴を計算する。次に、特徴を選択して、最終的に選択された特徴を取得する。具体的には、次の表1に示すように、最終的に選択された特徴数の表である。

【0173】

表1 特徴数表

【0174】

【表1】

データセット	サンプル数	非物理的モデル特徴	物理及び経験的ポテンシャルに基づく特徴	特徴総数
Platinum	484	129	19	148
TKI	144	129	19	148

ここで、訓練によって得られた目標予測モデルを対比してテストした結果を図15に示す。この図15には、実験的に測定され予測された $\Delta\Delta G$ 値との散布図が示されており、 $\Delta\Delta G$ とは、リガンドと受容体の結合自由エネルギーの相対差、即ち、変異前後のタンパク質がそれぞれの化合物に結合したときの対応する結合自由エネルギーの差である。ここで、図15の1番目の行は、結合自由エネルギーの相対差を予測するために非物理モデル特徴のみを使用した結果の概略図であり、図15の2番目の行は、結合自由エネルギーの相対差を予測するために、非物理モデル特徴及び物理的及び経験的ポテンシャル特徴の両方を使用した結果の概略図である。1番目の列は、従来技術1を用いてテストして得た結合自由エネルギーの相対差値の散布図である。2番目の列は、従来技術2を用いてテストして得た結合自由エネルギーの相対差値の散布図である。3番目の列は、本発明の技術を用いてテストして得た結合自由エネルギーの相対差値の散布図である。ここで、RMSE（平均二乗誤差）、Pearson（Pearson Correlation Coefficientは、2つのデータセットが1つのライン上にあるか否かを測定するものであり、変数間の線形関係を測定するものである）及びAUPRC（曲線の下面積減少の精度のリコール曲線）を評価指標として用いた。ここで、RMSE、Pearson及びAUPRCの各指標の平均値、最小値、最大値をそれぞれ計算した結果を表2に示す。

【0175】

表2 評価指標表

【0176】

【表 2】

特徴	方法	RMSE (kcal/mol)	Pearson	AUPRC
非物理的 モデル 特徴	従来技術1	0.87 (0.85, 0.90)	0.24(0.20, 0.27)	0.29(0.26, 0.33)
	従来技術2	0.76 (0.75, 0.77)	0.47(0.38, 0.50)	0.43(0.33, 0.50)
	本発明	0.75 (0.73, 0.76)	0.46(0.40, 0.49)	0.47(0.38, 0.48)
全ての 特徴	従来技術1	0.85 (0.84, 0.88)	0.25(0.21, 0.29)	0.31(0.26, 0.33)
	従来技術2	0.75 (0.75, 0.77)	0.50(0.38, 0.54)	0.48(0.34, 0.52)
	本発明	0.73 (0.72, 0.74)	0.54(0.47, 0.56)	0.50(0.43, 0.52)

ここで、全ての特徴について本発明のRMSE（小さければ小さいほどいい）指標の平均値は0.73であり、最小値は0.72であり、最大値は0.74であり、他の特徴について明らかに等分誤差が小さいことが分かる。本発明におけるPearson（大きければ大きいほどいい）指標は、他の先行技術よりも明らかに優れている。本発明におけるAUPRC指標は、他の従来技術よりも優れている。そのため、本発明では、従来の技術に対して、予測の正確性が明らかに向上する。さらに、図16に、対比試験結果におけるAUPRC指標の模式図を示す。ここで、各曲線の左から右への最初の丸は、 $\Delta\Delta G > 1.36 \text{ kcal/mol}$ を閾値とした場合、試験サンプルが対応する薬剤耐性結果を得た際に、薬剤耐性結果の対応を予測する精度と再現率を示す。各曲線の左から2番目の丸は、上位15% $\Delta\Delta G$ のテストサンプルを耐性判定結果とした場合の、予測すべき耐性判定結果に対応する正確性と再現率を示している。これらから明らかなように、本発明の技術は明らかに耐性があるか否かを区別するための性能を向上させることができる。

【0177】

なお、図2～図14のフローチャートの各ステップは、矢印によって示された順序で示されているが、必ずしも矢印によって示された順序で実行されるわけではない。これらのステップの実行は、本明細書に明示的に記載されていない限り、厳密な順序に限定されず、他の順序で実行されてもよい。さらに、図2～図14のステップの少なくとも一部は、必ずしも同じ時点で実行されるのではなく、異なる時点で実行されてもよい複数のステップ又は複数のステージを含んでもよく、これらのステップ又はステージの実行順序は必ずしも連続的ではなく、他のステップ又は他のステップのステップ又はステージの少なくとも一部と交互に又は交互に実行されてもよい。

【0178】

1つの実施例では、図17に示すように、予測モデルの訓練装置1700を提供する。この予測モデルの訓練装置1700は、ソフトウェアモジュール又はハードウェアモジュールを使用し、或いはこれらの組み合わせを使用してコンピュータ機器の一部としてもよい。この装置は、具体的には、サンプル取得モジュール1702、サンプル決定モジュール1704、訓練モジュール1706及び反復モジュール1708を含む。

【0179】

サンプル取得モジュール1702は、訓練サンプルセットを取得する。訓練サンプルセットは、各訓練サンプル、各訓練サンプルに対応する訓練サンプル重み、及び各訓練サンプルに対応する目標エネルギー特徴を含み、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、目標エネルギー特徴は、野生型エネルギー特徴及び変異型エネルギー特徴に基づいて取得され、野生型エネルギー特徴は、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得され、変異型エネルギー特徴は、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行うことによって取得される。

【0180】

サンプル決定モジュール1704は、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定する。

【0181】

訓練モジュール1706は、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得する。

【0182】

反復モジュール1708は、モデル訓練が完了するまで、基礎予測モデルに基づいて各訓練サンプルに対応する訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行し、目標予測モデルを取得する。目標予測モデルは、入力されたタンパク質情報及び入力された化合物情報に対応する相互作用状態情報を予測するために使用される。

【0183】

1つの実施例では、予測モデルの訓練装置1700は、以下のモジュールをさらに含む。

【0184】

事前訓練モジュールは、各訓練サンプル及び各訓練サンプルに対応する相互作用状態ラベルを取得し、訓練サンプルセットは、野生型タンパク質情報、変異型タンパク質情報、及び化合物情報を含み、野生型タンパク質情報及び化合物情報に基づいて結合初期エネルギー特徴抽出を行い、野生型初期エネルギー特徴を取得し、変異型タンパク質情報及び化合物情報に基づいて結合初期エネルギー特徴抽出を行い、変異型初期エネルギー特徴を取得し、野生型初期エネルギー特徴及び変異型初期エネルギー特徴に基づいて各訓練サンプルに対応する目標初期エネルギー特徴を決定し、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力して予測を行い、各訓練サンプルに対応する初期相互作用状態情報を取得し、初期予測モデルは、ランダムフォレストアルゴリズムを用いて構築され、各訓練サンプルに対応する初期相互作用状態情報及び各訓練サンプルに対応する相互作用状態ラベルに基づいて損失計算を行い、各訓練サンプルに対応する初期損失情報を取得し、事前訓練が完了するまで、初期損失情報に基づいて初期予測モデルを更新し、各訓練サンプルに対応する目標エネルギー特徴を初期予測モデルに入力して予測するステップを再度実行し、事前訓練予測モデル及び目標初期エネルギー特徴に対応する特徴重要度を取得し、事前訓練の完了時の前記各訓練サンプルに対応する損失情報に基づいて各訓練サンプルに対応する訓練サンプル重みを決定し、特徴重要度に基づいて目標初期エネルギー特徴から目標エネルギー特徴を選択する。

【0185】

1つの実施例では、事前訓練モジュールは、各訓練サンプルに対応する目標初期エネルギー特徴を初期予測モデルに入力する。初期予測モデルは、各訓練サンプルに対応する目標初期エネルギー特徴を現在の分割すべきセットとし、目標初期エネルギー特徴に対応する初期特徴重要度を計算し、初期特徴重要度に基づいて目標初期エネルギー特徴から初期分割特徴を決定し、初期分割特徴に基づいて各訓練サンプルに対応する目標初期エネルギー特徴を分割し、各分割サンプルに対応する目標初期エネルギー特徴を含む各分割結果を取得し、分割が完了するまで、各分割結果を現在の分割すべきセットとし、目標初期エネルギー特徴に対応する初期特徴重要度を計算するステップを再度反復的に実行し、各訓練サンプルに対応する初期相互作用状態情報を取得する。

【0186】

1つの実施例では、サンプル取得モジュール1702は、各訓練サンプルに対応する信頼度を取得し、信頼度に基づいて各訓練サンプルに対応する訓練サンプル重みを決定する。

【0187】

1つの実施例では、サンプル取得モジュール1702は、野生型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行い、野生型エネルギー特徴を取得し、変異型タンパク質情報及び化合物情報に基づいて結合エネルギー特徴抽出を行い、変異型エ

10

20

30

40

50

エネルギー特徴を取得し、野生型エネルギー特徴と変異型エネルギー特徴との差を計算し、目標エネルギー特徴を取得する。

【0188】

1つの実施例では、野生型エネルギー特徴は、第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴を含む。サンプル取得モジュール1702は、野生型タンパク質情報及び化合物情報に基づいて非物理的なスコアリング関数を用いて結合エネルギー特徴抽出を行い、第1の野生型エネルギー特徴を取得し、野生型タンパク質情報及び化合物情報に基づいて物理的な関数を用いて結合エネルギー特徴抽出を行い、第2の野生型エネルギー特徴を取得し、第1の野生型エネルギー特徴及び第2の野生型エネルギー特徴に基づいて融合を行い、野生型エネルギー特徴を取得する。

10

【0189】

1つの実施例では、変異型エネルギー特徴は、第1の変異型エネルギー特徴及び第2の変異型エネルギー特徴を含む。サンプル取得モジュール1702は、変異型タンパク質情報及び化合物情報に基づいて非物理的な関数を用いて結合エネルギー特徴抽出を行い、第1の変異型エネルギー特徴を取得し、変異型タンパク質情報及び化合物情報に基づいて物理的な関数を用いて結合エネルギー特徴抽出を行い、第2の変異型エネルギー特徴を取得し、第1の変異型エネルギー特徴及び第2の変異型エネルギー特徴に基づいて融合を行い、変異型エネルギー特徴を取得する。

【0190】

1つの実施例では、サンプル決定モジュール1704は、タンパク質ファミリー情報を取得し、タンパク質ファミリー情報に基づいて訓練サンプルセットを分割し、各訓練サンプルグループを取得し、訓練サンプル重みに基づいて各訓練サンプルグループから現在の訓練サンプルを選択し、現在の訓練サンプルセットを取得する。

20

【0191】

訓練モジュール1706は、現在の訓練サンプルセットにおける各現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、目標基礎予測モデルを取得する。

【0192】

1つの実施例では、サンプル決定モジュール1704は、現在の学習パラメータを取得し、現在の学習パラメータに基づいて選択サンプル数及びサンプル分布を決定し、選択サンプル数及びサンプル分布に基づいて訓練サンプル重みに従って各訓練サンプルグループから現在の訓練サンプルを選択し、目標の現在の訓練サンプルセットを取得する。

30

【0193】

1つの実施例では、訓練モジュール1706は、現在の訓練サンプルに対応する現在の目標エネルギー特徴を前記事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得し、現在の相互作用状態情報と現在の訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、現在の損失情報を取得し、基礎訓練の完了条件に達するまで、現在の損失情報に基づいて事前訓練予測モデルを更新し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して予測し、現在の相互作用状態情報を取得するステップを再度実行し、基礎予測モデルを取得する。

40

【0194】

1つの実施例では、反復モジュール1708は、各訓練サンプルに対応する目標エネルギー特徴を基礎予測モデルに入力し、各訓練サンプルに対応する基礎相互作用状態情報を取得し、各訓練サンプルに対応する基礎相互作用状態情報と各訓練サンプルに対応する相互作用状態ラベルとの誤差を計算し、基礎損失情報を取得し、基礎損失情報に基づいて訓練サンプル重みを更新し、各訓練サンプルに対応する更新サンプル重みを取得する。

【0195】

1つの実施例では、反復モジュール1708は、現在の学習パラメータを取得し、現在の学習パラメータに基づいて更新閾値を計算し、更新閾値と各訓練サンプルに対応する基礎損失情報とを比較し、各訓練サンプルに対応する比較結果を取得し、各訓練サンプルに

50

対応する比較結果に基づいて各訓練サンプルに対応する更新サンプル重みを決定する。

【0196】

1つの実施例では、現在の学習パラメータは、多様性学習パラメータ及び難易度学習パラメータを含む。反復モジュール1708は、各訓練サンプルグループを取得し、各訓練サンプルグループから現在の訓練サンプルグループを決定し、現在の訓練サンプルグループに対応するサンプルランクを計算し、サンプルランクに基づいて加重値を計算し、加重値を用いて多様性学習パラメータに対して重み付けを行い、目標加重値を取得し、目標加重値と難易度学習パラメータとの和を計算し、更新閾値を取得する。

【0197】

1つの実施例では、反復モジュール1708は、現在の学習パラメータを取得し、所定の増加量に従って現在の学習パラメータを更新し、更新学習パラメータを取得し、更新学習パラメータを現在の学習パラメータとする。

10

【0198】

1つの実施例では、図18に示すように、データ予測装置1800を提供する。この装置は、ソフトウェアモジュール、ハードウェアモジュール、又はこれらの組み合わせを用いてコンピュータ機器の一部としてもよい。該装置は、具体的には、データ取得モジュール1802、特徴抽出モジュール1804、目標特徴決定モジュール1806及び予測モジュール1808を含む。

【0199】

データ取得モジュール1802は、予測すべきデータを取得する。予測すべきデータは、予測すべき野生型タンパク質情報、予測すべき変異型タンパク質情報、及び予測すべき化合物情報を含む。

20

【0200】

特徴抽出モジュール1804は、予測すべき野生型タンパク質情報及び予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき野生型エネルギー特徴を取得し、予測すべき変異型タンパク質情報及び予測すべき化合物情報に基づいて結合エネルギー特徴抽出を行い、予測すべき変異型エネルギー特徴を取得する。

【0201】

目標特徴決定モジュール1806は、予測すべき野生型エネルギー特徴及び予測すべき変異型エネルギー特徴に基づいて予測すべき目標エネルギー特徴を決定する。

30

【0202】

予測モジュール1808は、予測すべき目標エネルギー特徴を目標予測モデルに入力して予測を行い、相互作用状態情報を取得する。目標予測モデルは、訓練サンプルセットを取得し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定し、現在の訓練サンプルに対応する現在の目標エネルギー特徴を事前訓練予測モデルに入力して基礎訓練を行い、基礎訓練が完了する際に、基礎予測モデルを取得し、モデル訓練が完了するまで、基礎予測モデルに基づいて訓練サンプル重みを更新し、訓練サンプル重みに基づいて訓練サンプルセットから現在の訓練サンプルを決定するステップを再度実行することによって取得される。

【0203】

予測モデルの訓練装置及びデータ予測装置の具体的な限定は、予測モデルの訓練方法及びデータ予測方法に関する上記の限定を参照することができ、ここでは説明しない。上記のデータ予測装置における各モジュールは、ソフトウェア、ハードウェア、及びそれらの組み合わせによって、全て又は一部が実現されてもよい。上記の各モジュールは、コンピュータ機器内のプロセッサにハードウェア的に組み込まれていても、プロセッサが上記の各モジュールに対応する動作を実行するために呼び出すことを容易にするために、コンピュータ機器内のメモリにソフトウェア的に格納されていてもよい。

40

【0204】

1つの実施例では、コンピュータ機器は、図19に示すような内部構成図を有するサーバとすることができる。コンピュータ機器は、システムバスによって接続するプロセッサ

50

、メモリとネットワークインタフェースを含む。ここで、コンピュータ機器のプロセッサは、計算と制御能力を提供するのに用いる。コンピュータ機器のメモリは、不揮発性記憶媒体、メモリを含む。不揮発性記憶媒体は、オペレーティングシステム、コンピュータ可読命令とデータベースを記憶する。メモリは、不揮発性記憶媒体中の操作システムとコンピュータ可読命令の運行に環境を提供する。このコンピュータ機器のデータベースは、訓練サンプルデータ及び予測すべきデータを格納するために使用される。このコンピュータ機器のネットワークインタフェースは、外部の端末とネットワーク接続で通信するためのものである。コンピュータ可読命令は、予測モデルの訓練方法又はデータ予測方法を実現するためにプロセッサによって実行される。

【0205】

1つの実施例では、コンピュータ機器を提供する。このコンピュータ機器は、図20に示すような内部構造図を有する端末とすることができる。このコンピュータ機器は、システムバスを介して接続されたプロセッサ、メモリ、通信インターフェース、ディスプレイ、及び入力装置を含む。ここで、コンピュータ機器のプロセッサは、計算と制御能力を提供するのに用いる。コンピュータ機器のメモリは、不揮発性記憶媒体、メモリを含む。この不揮発性記憶媒体は、オペレーティングシステム及びコンピュータ可読命令を記憶する。メモリは、不揮発性記憶媒体中の操作システムとコンピュータ可読命令の運行に環境を提供する。このコンピュータ機器の通信インターフェースは、WIFI、通信事業者ネットワーク、NFC（近距離通信）、又は他の技術によって実現される外部の端末と有線又は無線で通信するために使用される。コンピュータ可読命令はプロセッサに実行される時、予測モデルの訓練方法とデータ予測方法を実現する。コンピュータ機器の表示装置は、液晶表示装置又は電子インク表示装置である、コンピュータ機器の入力装置は、表示装置に覆うタッチ層である、コンピュータ機器のケーシングに設置するキー、トラックボール又はトラックパッドであってもよいし、外付けのキーボード、トラックパッド又はマウスなどであってもよい。

【0206】

当業者であれば、図19及び図20に示された構造は、本発明の技術に関連する部分的な構造のブロック図にすぎず、本発明の技術が適用されるコンピュータ機器の限定を構成しないこと、特定のコンピュータ機器は、図示されたものよりも多くの又は少ない構成要素を含むことができること、ある構成要素を組み合わせることができること、又は異なる構成要素の配置を有することができることを理解するであろう。

【0207】

1つの実施例では、コンピュータ可読命令が格納されたメモリと、コンピュータ可読命令を実行するときに上述の方法の実施形態のステップを実施するプロセッサとを含むコンピュータ機器も提供される。

【0208】

1つの実施例では、上述の方法の実施形態のステップを実行するためにプロセッサによって実行されるコンピュータ可読命令を格納するコンピュータ可読記憶媒体が提供される。

【0209】

1つの実施例では、コンピュータ可読記憶媒体に記憶されたコンピュータ命令を含むコンピュータプログラム製品又はコンピュータプログラムが提供される。コンピュータ機器のプロセッサは、コンピュータ可読記憶媒体からコンピュータ命令を読み取り、プロセッサは、コンピュータ機器が上述の方法の様々な実施形態のステップを実行するようにコンピュータ命令を実行する。

【0210】

当業者であれば、上述した実施形態の方法のプロセスの全部又は一部を実現することは、実行時に上述した方法の実施形態のようなプロセスを含むことができる不揮発性のコンピュータ可読記憶媒体に格納され得るコンピュータプログラムによって関連するハードウェアを命令することによって達成され得ることを理解するであろう。ここで、本発明に提供される実施形態で使用されるメモリ、ストレージ、データベース、又は他の媒体への任

意の参照は、不揮発性メモリ及び揮発性メモリのうちの少なくとも1つを含むことができる。不揮発性メモリは、リードオンリーメモリ（Read-Only Memory：ROM）、磁気テープ、フロッピーディスク、フラッシュメモリ、又は光メモリなどを含むことができる。揮発性メモリは、ランダムアクセスメモリ（Random Access Memory：RAM）又は外部キャッシュを含むことができる。限定ではなく例示として、RAMは、スタティックランダムアクセスメモリ（Static Random Access Memory：SRAM）又はダイナミックランダムアクセスメモリ（Dynamic Random Access Memory：DRAM）などの様々な形態であってもよい。

【0211】

以上の実施例の各技術的特徴は、任意の組み合わせが可能であり、説明を簡潔にするために、上記実施例に係る各技術的特徴の可能な全ての組み合わせについては説明しなかったが、これらの技術的特徴の組み合わせに矛盾がない限り、本明細書に記載された範囲と考えるべきである。

【0212】

上記の実施形態は、本出願の幾つかの実施形態のみを示しており、その説明はより具体的で詳細であるが、従って、発明の特許範囲を限定するものとして理解されるべきではない。当業者であれば、本出願の概念から逸脱することなく、本出願の範囲内で幾つかの変形及び修正を行うことができることに留意されたい。従って、本出願の特許の保護範囲は添付の特許請求の範囲により決まる。

10

20

30

40

50

【図面】
【図 1】

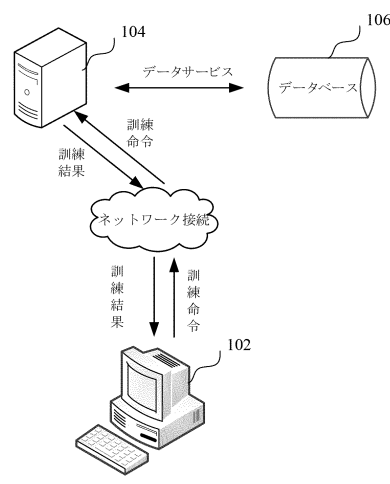


図 1

【図 2】

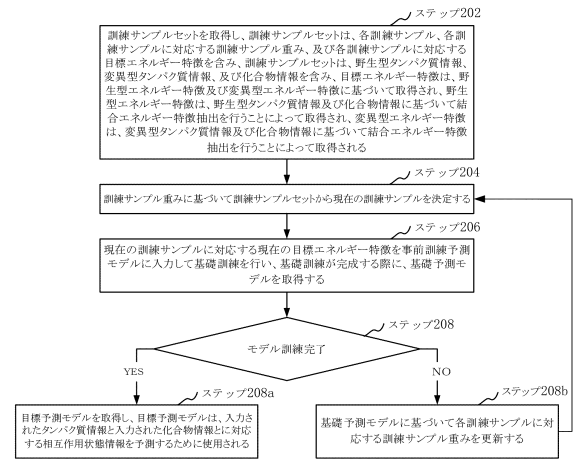


図 2

【図 3】

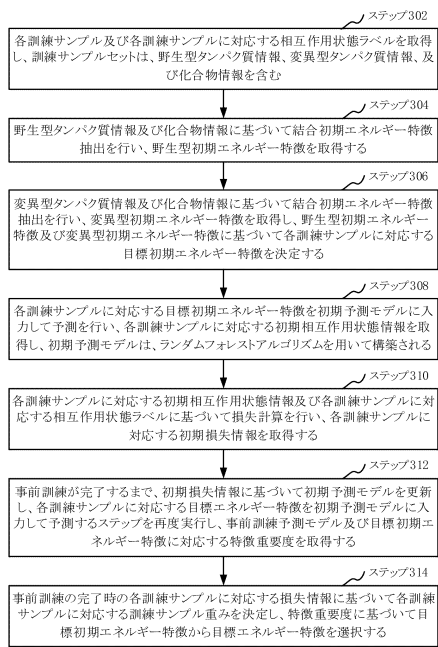


図 3

【図 4】

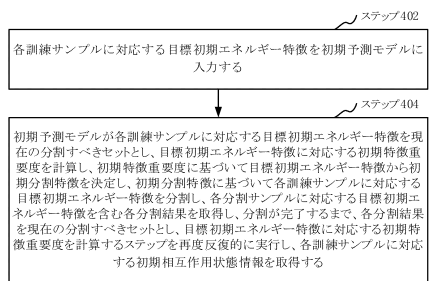


図 4

【図 5】

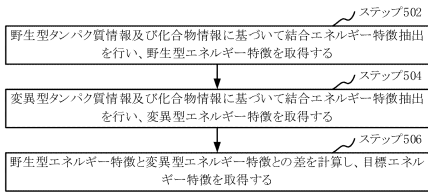


図 5

【図 6】

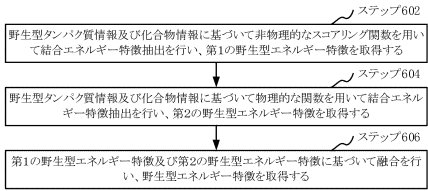


図 6

【図 7】

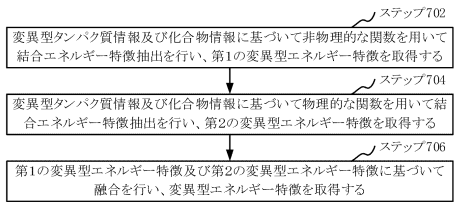


図 7

【図 8】

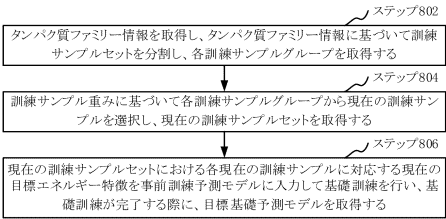


図 8

【図 9】

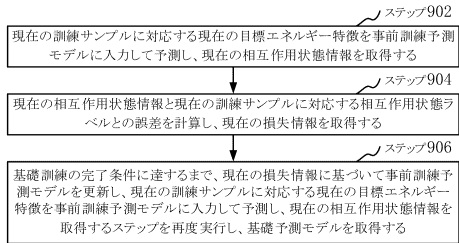


図 9

【図 10】

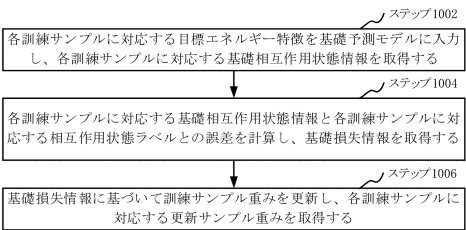


図 10

10

20

30

40

50

【図 1 1】

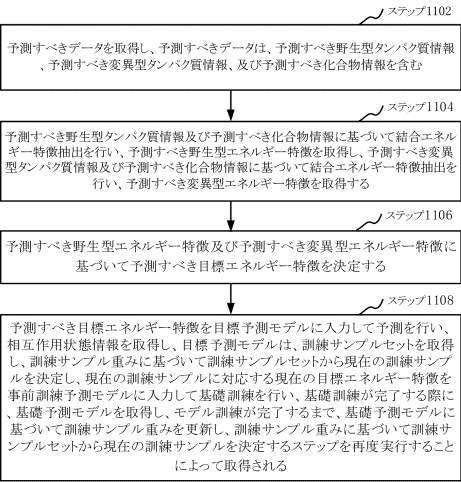


図 11

【図 1 2】

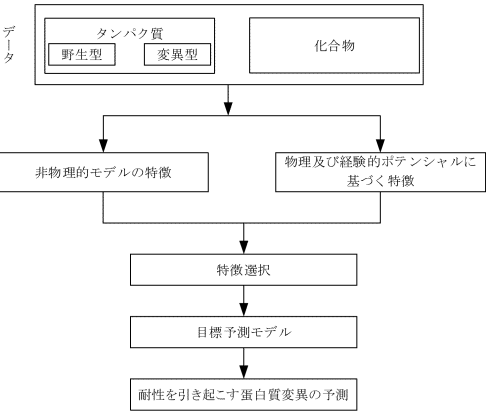


図 12

10

【図 1 3】

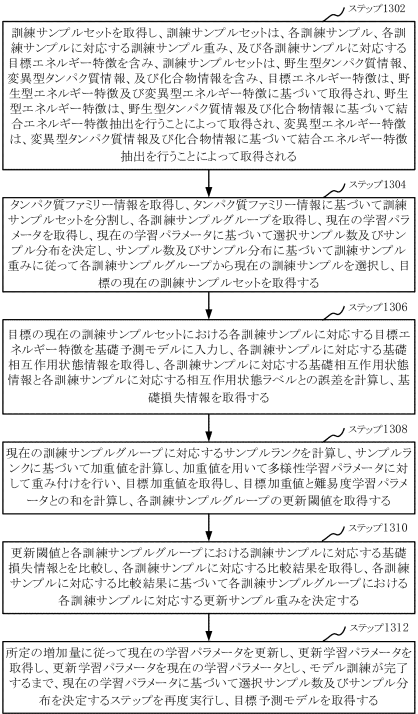


図 13

【図 1 4】

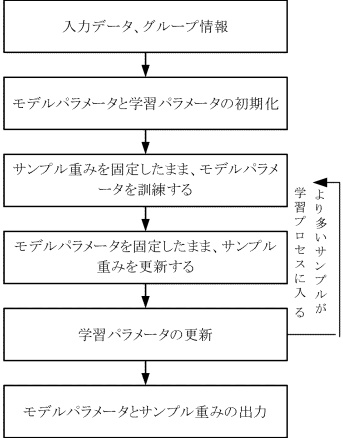


図 14

20

30

40

50

【図 15】

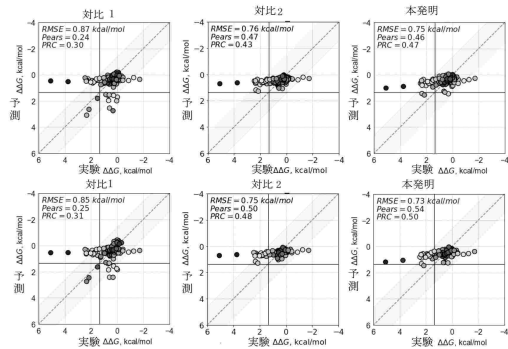


图 15

【図 16】

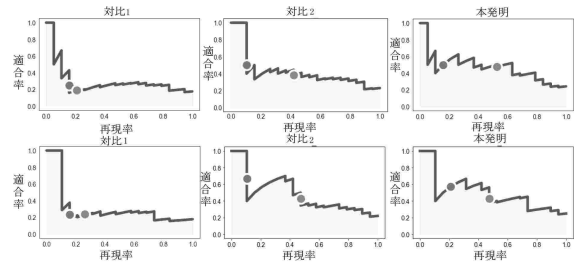


图 16

10

【図 17】

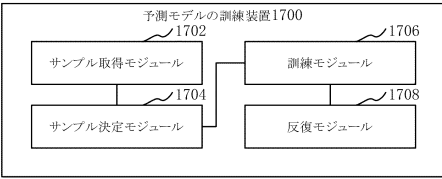


图 17

【図 18】

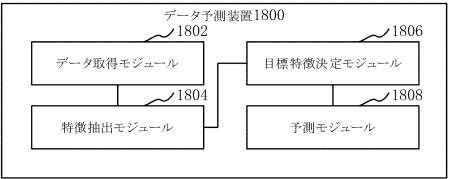


图 18

20

【図 19】

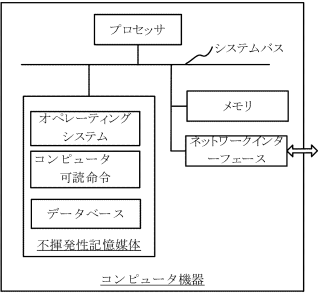


图 19

【図 20】

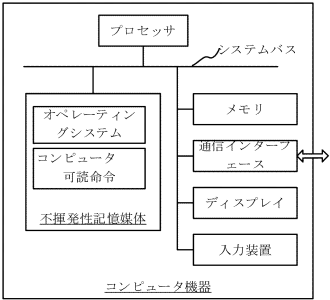


图 20

30

40

50

フロントページの続き

- (72)発明者 ヤン, ツィイー
中華人民共和国 518057 グアンドン シェンジェン ナンシャン・ディストリクト ミッドウ
エスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディ
ング 35エフ
- (72)発明者 イエ, ジャオフェン
中華人民共和国 518057 グアンドン シェンジェン ナンシャン・ディストリクト ミッドウ
エスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディ
ング 35エフ
- (72)発明者 リアオ, ベンベン
中華人民共和国 518057 グアンドン シェンジェン ナンシャン・ディストリクト ミッドウ
エスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディ
ング 35エフ
- (72)発明者 ジャン, シェンユー
中華人民共和国 518057 グアンドン シェンジェン ナンシャン・ディストリクト ミッドウ
エスト・ディストリクト・オブ・ハイテックパーク ケジジョンギ・ロード テンセント・ビルディ
ング 35エフ
- 審査官 渡邊 加寿磨
- (56)参考文献 中国特許出願公開第112289371 (CN, A)
中国特許出願公開第109147866 (CN, A)
米国特許出願公開第2019/0304568 (US, A1)
米国特許出願公開第2007/0192037 (US, A1)
特開2017-189176 (JP, A)
- (58)調査した分野 (Int.Cl., DB名)
G06Q 10/00-99/00
G16B 5/00-99/00
G16Z 99/00
G06N 20/00